DOCTORAL DISSERTATION

Applications of machine learning algorithms in quality of life

---

By:

Christos Kokkotis

Supervisor:

Dr. Dimitrios E. Tsaopoulos

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in the

Department of Physical Education & Sport Sciences,

University of Thessaly

Trikala, September 2021

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Members of the examination committee

Full name	University/Research centre	Information
Dr. Dimitrios E. Tsaopoulos	Researcher B', Bio-Economy and Agri-Technology (IBO)/ Centre for Research and Technology – Hellas (CERTH)	Supervisor
Dr. Giannis Giakas	Professor, Department of Physical Education & Sport Science, University of Thessaly, Greece	Three-member committee
Dr. Elpiniki Papageorgiou	Associate Professor, Department of Energy Systems, University of Thessaly, Greece	Three-member committee
Dr. Athanasios Tsiokanos	Professor, Department of Physical Education & Sport Science, University of Thessaly, Greece	Seven-member committee
Dr. Dimitrios Paraskevis	Associate Professor, Department of Hygiene and Epidemiology, Medical School, National and Kapodistrian University of Athens, Greece	Seven-member committee
Dr. Sotirios Tasoulis	Assistant Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece	Seven-member committee
Dr. Georgios Banias	Researcher B', Bio-Economy and Agri-Technology (IBO)/ Centre for Research and Technology – Hellas (CERTH)	Seven-member committee

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Declaration

Christos Kokkotis, hereby declare that this thesis has not been previously submitted in this University or any other University for the award of any degree, diploma, associateship, fellowship or other similar titles of recognition.

Thesis title: Applications of machine learning algorithms in quality of life,

Εφαρμογές αλγορίθμων μηχανικής μάθησης στην ποιότητας ζωής

Pages: 175

Words: 47896

Figures: 42

Tables: 32

Full name: Christos I. Kokkotis
Signed by: Christos I. Kokkotis
Submission date: September 02, 2021
Revision date: October 06, 2021

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Plagiarism Disclaimer

I declare that this thesis is based on a total of five studies/papers of which three (Chapter 1, Chapter 2, and Chapter 3) have been published and two have been submitted for publication. These papers are referred to and cited in the thesis chapters (with specification in the intro to each chapter of the study/paper referenced).

Published articles

Chapter 1 Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. E. (2020). Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open, 2(3), 100069.

Chapter 2 Kokkotis, C., Moustakidis, S., Giakas, G., & Tsaopoulos, D. (2020). Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients. Applied Sciences, 10(19), 6797.

Chapter 3 Kokkotis, C, Moustakidis, S, Baltzopoulos, V, Giakas, G, & Tsaopoulos, D (2021) Identifying Robust Risk Factors for Knee Osteoarthritis Progression: An Evolutionary Machine Learning Approach. Healthcare 9(3) 260.

Articles under review (Unpublished data)

Chapter 4 Explainable Machine Learning for Knee Osteoarthritis Diagnosis Based on a Novel Fuzzy Feature Selection Methodology

Chapter 5 Leveraging explainable machine learning to identify gait biomechanical parameters associated with Anterior Cruciate Ligament injury

---

Full name: Christos I. Kokkotis
Signed by: Christos I. Kokkotis
Submission date: September 02, 2021
Revision date: October 06, 2021

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Acknowledgments

The joy of completing this PhD thesis is not easily described in words. The only thing I can say for sure is that I feel blessed for the friends I gained through this journey.

First and foremost, I would like to express my sincere gratitude to my supervisor and friend, Dr. Dimitrios E. Tsaopoulos for his guidance during the research and writing of this thesis, for his patience and his continuous support in whatever came up, both at work and in my personal life. My sincere gratitude also goes to Dr. Giannis Giakas, for giving me the opportunity to conduct this PhD program, for his constant support, guidance, and his constructive suggestions.

A special thanks to my mentor and friend, Dr. Serafeim Moustakidis. Words are too few to express my gratitude for his continuous support, patience, and immense knowledge. During the whole time of this journey his guidance helped me to overcome any challenging situation. Furthermore, I would like to thank, my friend Dr. Themistoklis Tsatalas for his help at any time, availability and for his continuous support. Moreover, I would like to thank my friends Dr. Charis Ntakolia, Dimitrios Tikas as well as the personnel of the ErgoMech-Lab at the University of Thessaly for their valuable help.

This work was supported by funding from the European Community's H2020 Programme, under grant agreement Nr. 777159 (OActive project) and by the National Operational Program Competitiveness, Entrepreneurship and Innovation, under the call Research – Create – Innovate (SafeACL project, code: T1EDK-04234).

Last but not least, I would like to especially thank my family and friends for their understanding, never-ending encouragement and support throughout this journey, because without them I would not have reached my goal.

Christos Kokkotis

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Table of Contents

List of Figures 10
List of Tables 14
Abstract 17
Περίληψη 19
General Introduction 21
Chapter 1 25
Abstract 25
Introduction 26
Machine learning in a nutshell 27
Methods 30
Literature Search Approach 30
Exclusion Criteria 31
Assessed Outcomes 31
Results 31
Predictions/Regression 31
Classification 34
Biomechanical data and discrete variables 34
Medical Images 37
Optimum post-treatment planning techniques 43
Segmentation 44
Discussion and Conclusions 47
Declaration of Competing Interest 50
Acknowledgments 50
Chapter 2 51
Abstract 51
Introduction 52
Data Description 53
Methodology 61
Pre-Processing 62
Feature Selection (FS) 62
Learning Process 63

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Validation 65
Results 66
Prediction Performance 66
Comparative Analysis 78
Discussion 79
Conclusions 82
Conflicts of Interest 82
Data Availability Statement 82
Funding 82
Chapter 3 84
Abstract 84
Introduction 85
Methods 87
Dataset Description 87
Problem Definition 88
Data Pre-Processing 89
Feature Selection 89
Learning 93
Validation 94
Explainability 94
Results 95
Selection Criterion 95
Features Selected 98
Comparative Analysis 99
Explainability Results 103
Discussion 104
Conclusions 107
Conflicts of Interest 107
Data Availability Statement 107
Funding 108
Chapter 4 109
Abstract 109
Introduction 110

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Materials and Methods 112
Dataset Description 112
Methodology 113
Problem Definition 113
Data Pre-processing 113
Proposed FS methodology 114
Learning 115
Validation 115
Explainability 116
Results and Discussion 116
A. Results 116
Diagnosis Performance 116
Features Selected 117
Comparative Analysis 118
Explainability Results 119
B. Discussion 121
Conclusions 122
Conflicts of Interest 122
Data Availability Statement 122
Funding 123
Chapter 5 124
Abstract 124
Introduction 125
Materials and Methods 126
Participants 126
Testing procedure and data collection 127
Data Analysis 128
Machine Learning workflow 130
Statistical Analysis 131
Results 132
Comparative Analysis 132
Explainability Results 134
Global exploration 134

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Local exploration 134
Statistical Analysis 136
Discussion of Results 138
Conclusions 140
Conflicts of Interest 141
Data Availability Statement 141
Funding 141
Institutional Review Board Statement 141
Informed Consent Statement 141
General Conclusions 142
References 145
Appendixes 162
Appendix A 162
Appendix B 165
Appendix C 167
Annexes 169
Annex A: Ethics 169
Annex B: Candidate's responsibilities throughout the study 170
Annex C: Skills acquired during the PhD programme 171
Annex D: Publications during PhD Studies 172
Annex E: Awards during PhD Studies 174

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

List of Figures

#	Title	Page
1	Musculoskeletal Disorders	21
2	Kellgren-Lawrence (KL) scale	22
3	Anterior cruciate ligament (ACL) injury	23
1.1	A typical machine learning system	28
1.2	a) A temporal evolution chart depicting the number of papers per category published each year since the year 2006 and included in the survey, b) Bubble chart showing a distribution of the papers considered in this survey arranged according to the data sources utilized in each survey category	48
2.1	Flow chart of study design for dataset A	56
2.2	Flow chart of study design for dataset B	58
2.3	Flow chart of study design for dataset C	59
2.4	Flow chart of study design for dataset D	60
2.5	Flow chart of study design for dataset E	61
2.6	Pseudocode for the implementation of the proposed feature selection (FS)	63
2.7	Learning curves with testing accuracy scores on dataset A for different machine learning (ML) models trained on feature subsets of increasing dimensionality	67
2.8	Learning curves with testing accuracy scores on dataset B for different ML models trained on feature subsets of increasing dimensionality	69
2.9	Learning curves with testing accuracy scores on dataset C for different ML models trained on feature subsets of increasing dimensionality	71
2.10	Learning curves with testing accuracy scores on dataset D for different ML models trained on feature subsets of increasing dimensionality	74

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

2.11	Learning curves with testing accuracy scores on dataset E for different ML models trained on feature subsets of increasing dimensionality	75
2.12	Features selected in datasets A to E in (**a**–**e**), respectively. Axis y (selection criterion) denotes how many times a feature has been selected (6 declares that a specific feature has been selected by all six FS techniques and so on). Features have been ranked based on the selection criterion V_j and are visualised with different colors each one representing a specific feature category	78
3.1	Stratification of the patients in our study and formulation of the training dataset. Inclu-sion/exclusion criteria are presented along with the definition of the two data classes (knee os-teoarthritis (KOA) progressors and non-progressors)	89
3.2	The proposed GenWrapper feature selection (FS) methodology that includes all the involved processing steps: (i) generation of the initial population; (ii) fitness measurement approach; (iii) stopping criterion; (iv) evolution mechanisms and (v) final feature ranking after the termination of the genetic algorithm (GA).	90
3.3	Definition of genes, chromosomes and population	92
3.4	Proposed mechanism for estimating the fitness of each chromosome within a generation	93
3.5	Fitness with respect to number of generations for GenWrapper. The black and blue dashed lines show the best and the mean fitness achieved at each generation, respectively	96
3.6	Feature ranking produced by the proposed FS (the dashed line indicates the number of features that were finally selected)	97
3.7	Accuracy (mean 10-fold cross-validation (10FCV)) with respect to selected features (curves): GenWrapper versus a classical wrapper using two classifiers (support vector machine (SVM) and logistic regression (LR))	100
3.8	Accuracy (mean 10FCV) with respect to selected features: GenWrapper versus the remaining competing FS techniques. SVM was used for the classification task for all eight FS techniques	101

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

3.9	Bar graph comparison for the best models (SVMs trained on the optimum number of selected features per case). Red lines correspond to the mean 10FCV, blue boxes visualize the standard deviation of the obtained accuracies, dashed black lines show the min–max range and the red crosses depict outliers (if any)	103
3.10	This figure depicts: (a) the SHAP summary plot and; (b) the SHAP feature importance for the SVM trained on the features selected by the proposed GenWrapper	104
4.1	The proposed AI methodology for KOA diagnosis	113
4.2	Feature Selection method based on Fuzzy Logic flowchart	114
4.3	Fuzzy set of input variables for FIS 1 and 2	115
4.4	Fuzzy set of output variable for FIS 1 and 2.	115
4.5	Curves with testing accuracy scores with respect to the number of selected features for different ML models	117
4.6	The 21 most informative selected risk factors per category	118
4.7	a) Features' impact on Random Forest (21F) model output for the testing set of OAI dataset. b) Features' average impact magnitude for testing instances	120
4.8	Risk factors contributions to ML model output for a KOA status subject	120
5.1	Three dimensional GRFs (a), sagittal plane kinematic (b) and kinetic (c) variables of interest during walking	129
5.2	The proposed AI workflow for ACL diagnosis and interpretation	130
5.3	Learning curves with testing accuracy scores for different ML models trained on feature subsets of increasing dimensionality in the 3-class problem (refering to both ACL deficient and ACL reconstructed patients)	133
5.4	Average feature impact magnitude for all instances in the 3-class problem	134
5.5	Features' impact on SVM model output for local problem 1. This figure shows the average impact magnitude for all instances in	135

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

	the task of differentiating the control group vs pre-surgery group	
5.6	Average feature impact magnitude for all instances in the local problem 2 (control versus ACLR)	136
5.7	Average feature impact magnitude for all instances for local problem 3 (pre-surgery group versus post-surgery group)	136

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

List of Tables

#	Title	Page
1.1	Presentation of indicative ML models along their characteristics	29
1.2	Studies with Predictions/Regression techniques	32
1.3	Classification studies employing biomechanical data and/or distinct variables	35
1.4	Medical image-based classification studies of KOA	38
1.5	Studies with ML-driven post-treatment planning techniques of KOA	43
1.6	Segmentation techniques applied on the KOA research	45
2.1	Main categories of the feature subsets considered in this work	54
2.2	Hyperparameters description	64
2.3	Confusion matrix	66
2.4	Best testing accuracies achieved for ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. A1 and A2 denote classes 1 and 2 of dataset A, respectively	67
2.5	Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. B1 and B2 denote classes 1 and 2 of dataset B, respectively	70
2.6	Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. C1 and C2 denote classes 1 and 2 of dataset C, respectively	71
2.7	Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. D1 and D2 denote classes 1 and 2 of dataset D, respectively	74
2.8	Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the	75

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

	hyperparameters of the ML models employed. E1 and E2 denote classes 1 and 2 of dataset E, respectively	
2.9	Summary of all reported results	77
2.10	Testing performance (%) of the competing FS techniques with respect to the number of selected features for dataset D	79
3.1	Main categories of the feature subsets considered in this study. A brief description is given along with the number of features considered per category and for each of the two visits	88
3.2	Hyperparameters of the optimized GenWrapper algorithm. A brief description of each hyperparameter is provided along with the finally selected value	93
3.3	Comparative analysis with respect to the final selection of features: proposed feature ranking versus the feature subset of the best individual solution in the final generation	97
3.4	Characteristics of the 35 most informative risk factors as selected by the proposed GenWrapper	98
3.5	Best performance (mean 10FCV) achieved by all competing FS techniques employing SVM along with the number of selected features in which this accuracy was accomplished	101
4.1	Main categories of the clinical evaluation data considered in this study	112
4.2	Summary of best metrics per model and number of selected features	117
4.3	Comparative analysis of FS methods	119
5.1	Subjects' characteristics	127
5.2	Evaluated parameters of gait cycle for vertical and horizontal GRFs and the sagittal plane kinematics and kinetics	129
5.3	Best testing accuracies (%) achieved for ML models in 3-class problem along with precision, recall, f1-score and the optimum number of features	133
5.4	Statistical comparison at the global level	136

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

5.5	Statistical analysis at the local level for ACL diagnosis and postoperatively	137
A	Selected features that led to the overall best Knee Osteoarthritis (KOA) prediction performance in our study	162
B	Selected features that led to the best overall KOA prediction performance in our study. The features have been ranked according to their impact on the classification result as calculated by SHapley Additive exPlanations (SHAP).	165
C	The 21 most informative selected risk factors as described in OAI database	167

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Abstract

According to World Health Organization, a recent analysis showed that 1.71 billion people globally have musculoskeletal conditions. The societal impact in terms of direct healthcare costs and indirect (i.e., productivity loss) costs is enormous. Hence, it is vital to understand the pathophysiology of musculoskeletal diseases using artificial intelligence analytics tools with ultimate objective to develop techniques for their interpretation, diagnosis, prediction and rehabilitation. The aim of this thesis is to extend the current understanding of the contribution of the risk factors in the development of Knee Osteoarthritis and to uncover the rationale behind the biomechanical parameters from the anterior cruciate ligament post-surgery rehabilitation in order to avoid the outset of KOA. To achieve these goals, first of all we conducted a review about the machine learning techniques in knee osteoarthritis. Subsequently, we employed data from the osteoarthritis initiative (OAI) database (available on https://nda.nih.gov/oai/) and collected numerous biomechanical data from individuals who suffered from anterior cruciate ligament injury or not. This work led to five studies which are presented as different chapters of the current thesis. The review guided us to understand the literature gap and to develop machine learning techniques related to the prognosis and diagnosis of knee osteoarthritis as well as the interpretation of these models. In the second work, we used data from OAI and we worked on the prediction of KOA through the identification of risk factors that are relevant with KL progression. One of the main objectives of this work was to explore three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. The findings of this work were the input for the third work. So, the next step in the prediction task was to apply an evolutionary genetic algorithm (GA)-based wrapper technique, which leads to selected features that consistently work well at any possible data sample and, thus, have increased generalization capacity with respect to KOA progression. The impact of the selected risk factors on the prediction output was further investigated using SHapley Additive exPlanations (SHAP). The fourth work focused on the diagnosis task and interpretation of the model output. The objective of the present study was to provide a robust feature selection methodology based on fuzzy logic that could: (i) handle the multidimensional nature of the available datasets (OAI) and (ii) alleviate the defectiveness of existing feature selection techniques towards the identification of important risk factors which contribute to KOA diagnosis and interpretation. The fifth work has the aim to investigate the modification of the biomechanical parameters after an ACL injury, which is a risk factor for the onset of KOA. For this aim, a state-of-the-art explainability analysis based on SHAP and conventional statistical analysis attempted to uncover the rationale behind the decision-making mechanism of the best trained model and provide a holistic approach of quantifying the contribution of the

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

input gait biomechanical parameters in the tasks of ACL injury diagnosis. The proposed AI methodologies may contribute to the development of new, efficient risk stratification strategies and identification of risk phenotypes of each KOA patient to enable appropriate interventions. Furthermore, features, that would have been neglected by the traditional statistical analysis, were identified as contributing parameters having significant impact on the ML model's output for prediction of KOA progression, KOA diagnosis, ACL injury diagnosis during gait.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Περίληψη

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, μια πρόσφατη μελέτη έδειξε ότι 1,71 δισεκατομμύρια άνθρωποι παγκοσμίως πάσχουν από μυοσκελετικές παθήσεις. Ο κοινωνικός αντίκτυπος όσον αφορά το άμεσο κόστος υγειονομικής περίθαλψης αλλά και το έμμεσο (δηλ. απώλεια παραγωγικότητας) είναι τεράστιος. Ως εκ τούτου, είναι ζωτικής σημασίας η κατανόηση της παθοφυσιολογίας των μυοσκελετικών παθήσεων, έτσι ώστε με τη χρήση αναλυτικών εργαλείων τεχνητής νοημοσύνης να αναπτυχθούν τεχνικές για την ερμηνεία, τη διάγνωση, την πρόβλεψη και την αποκατάστασή τους. Σκοπός της παρούσας διδακτορικής διατριβής είναι να διευρύνει την τρέχουσα κατανόηση της συμβολής των παραγόντων κινδύνου στην ανάπτυξη της οστεοαρθρίτιδας γόνατος και να αποκαλύψει την επίδραση των εμβιομηχανικών παραμέτρων στη μετεγχειρητική αποκατάσταση του πρόσθιου χιαστού συνδέσμου, προκειμένου να αποφευχθεί η εμφάνιση οστεοαρθρίτιδας γόνατος. Για την επίτευξη των παραπάνω στόχων, αρχικά πραγματοποιήσαμε μια βιβλιογραφική ανασκόπηση σχετικά με τις τεχνικές μηχανικής μάθησης στην οστεοαρθρίτιδα γόνατος. Στη συνέχεια, χρησιμοποιήσαμε δεδομένα από τη βάση δεδομένων της πρωτοβουλίας για την οστεοαρθρίτιδα (OAI) (διαθέσιμα στη διεύθυνση https://nda.nih.gov/oai/) και συλλέξαμε επίσης πληθώρα εμβιομηχανικών δεδομένων από άτομα που υπέφεραν ή όχι από τραυματισμό πρόσθιου χιαστού συνδέσμου. Η προεργασία αυτή οδήγησε σε πέντε μελέτες, οι οποίες παρουσιάζονται ως διαφορετικά κεφάλαια της τρέχουσας διατριβής. Η βιβλιογραφική ανασκόπηση μας οδήγησε στο να κατανοήσουμε το κενό στη βιβλιογραφία και να αναπτύξουμε μια σειρά τεχνικών μηχανικής μάθησης που σχετίζονται με την πρόγνωση και τη διάγνωση της οστεοαρθρίτιδας γόνατος καθώς και την ερμηνεία των μοντέλων αυτών. Στη συνέχεια, στην δεύτερη μελέτη χρησιμοποιήσαμε δεδομένα από την βάση OAI και δουλέψαμε πάνω στην πρόβλεψη της οστεοαρθρίτιδας γόνατος, μέσω του εντοπισμού παραγόντων κινδύνου που σχετίζονται με την εξέλιξη του βαθμού KL. Στη συνέχεια, ο κύριος στόχος αυτής της εργασίας ήταν να διερευνηθούν τρεις διαφορετικές επιλογές όσον αφορά τη χρονική περίοδο εντός της οποίας θα πρέπει να ληφθούν υπόψη τα δεδομένα προκειμένου να προβλεφθεί αξιόπιστα η εξέλιξη της οστεοαρθρίτιδας γόνατος. Τα ευρήματα αυτής της εργασίας αποτέλεσαν τη πηγή δεδομένων για τη τρίτη μελέτη. Έτσι, το επόμενο βήμα για την πρόβλεψη της οστεοαρθρίτιδας γόνατος ήταν η εφαρμογή μιας εξελικτικής τεχνικής περιτύλιξης με βάση τον γενετικό αλγόριθμο, η οποία οδηγεί σε επιλεγμένα χαρακτηριστικά που λειτουργούν αξιόπιστα και αποδοτικά σε οποιοδήποτε πιθανό δείγμα δεδομένων και, συνεπώς, έχουν αυξημένη ικανότητα γενίκευσης σε σχέση με την πρόβλεψη της οστεοαρθρίτιδας γόνατος. Ο αντίκτυπος των επιλεγμένων παραγόντων κινδύνου στην διαμόρφωση της εξόδου του μοντέλου

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

πρόβλεψης, διερευνήθηκε περαιτέρω χρησιμοποιώντας το εργαλείο ερμηνείας SHAP. Η τέταρτη μελέτη επικεντρώθηκε στην διάγνωση της οστεοαρθρίτιδας γόνατος. Ο στόχος της παρούσας μελέτης ήταν να παράσχει μια ισχυρή μεθοδολογία επιλογής χαρακτηριστικών (FS) που θα μπορούσε: (i) να χειριστεί την πολυδιάστατη φύση των διαθέσιμων συνόλων δεδομένων (OAI) και (ii) να αντιμετωπίσει τα μειονεκτήματα των υφιστάμενων τεχνικών επιλογής χαρακτηριστικών για τον εντοπισμό σημαντικών παραγόντων κινδύνου που συμβάλλουν στη διάγνωση της οστεοαρθρίτιδας γόνατος αλλά και την ερμηνεία της. Η πέμπτη μελέτη έχει ως στόχο να διερευνήσει την προσαρμογή των εμβιομηχανικών παραμέτρων μετά από τραυματισμό του πρόσθιου χιαστού συνδέσμου, ο οποίος αποτελεί παράγοντα κινδύνου για την εμφάνιση οστεοαρθρίτιδας γόνατος. Για το σκοπό αυτό, μια καινοτόμος ανάλυση επεξηγήσεων βασισμένη στο εργαλείο SHAP και τη συμβατική στατιστική ανάλυση προσπάθησε να αποκαλύψει το σκεπτικό πίσω από τον μηχανισμό λήψης αποφάσεων του καλύτερα εκπαιδευμένου μοντέλου διάγνωσης και να παράσχει μια ολιστική προσέγγιση ποσοτικοποίησης της συμβολής των εμβιομηχανικών παραμέτρων βάδισης στις διεργασίες της διάγνωσης πρόσθιου χιαστού συνδέσμου και της μετεγχειρητικής αποκατάστασης αυτού. Οι προτεινόμενες μεθοδολογίες τεχνητής νοημοσύνης μπορούν να συμβάλουν στην ανάπτυξη νέων, αποτελεσματικών στρατηγικών διαστρωμάτωσης του κινδύνου και στον εντοπισμό παραγόντων κινδύνου εξατομικευμένα σε κάθε πάσχοντα από οστεοαρθρίτιδα γόνατος, ώστε να αναπτυχθούν εξατομικευμένες παρεμβάσεις. Επιπλέον, παράμετροι που δε θα είχαν αναδειχθεί από την παραδοσιακή στατιστική ανάλυση, προσδιορίστηκαν ως παράμετροι που έχουν σημαντικό αντίκτυπο στην έξοδο του μοντέλου μηχανικής μάθησης, τόσο για την πρόβλεψη της εξέλιξης της οστεοαρθρίτιδας γόνατος, όσο και για τη διάγνωση της οστεοαρθρίτιδας γόνατος, τη διάγνωση της ρήξης πρόσθιου χιαστού και τη μετεγχειρητική αποκατάσταση αυτού.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

General Introduction

Applications of machine learning algorithms in quality of life

Musculoskeletal disorders (MSDs) comprise more than 150 conditions that affect the human body's movement or musculoskeletal system (i.e., tendons, muscles, nerves, ligaments, blood vessels, discs, etc.). The symptoms of musculoskeletal conditions could include stiff joints, swelling and recurrent pain and they can affect major areas of our body (e.g., knees, hips and shoulders as shown in Figure 1). The main factors that cause musculoskeletal conditions are occupation, age, injuries, obesity, activity level, family history and lifestyle. Musculoskeletal conditions exist in numerous occupations and they are one of the most important occupational problems. Hence MSDs affect the quality of life, the personnel's health and job satisfaction. According to the existing literature, more than 30% of workers in Europe suffer from MSDs (almost 40 million). Osteoarthritis (OA), rheumatoid arthritis, tendinitis, fibromyalgia, carpal tunnel syndrome and bone fractures are the most common MSDs diseases.

Figure 1. Musculoskeletal Disorders.

The common characteristic of all the above diseases is the multifactorial causality. According to World Health Organization 343 million people globally suffer from OA [1]. Hence, the present PhD thesis focuses on the Knee Osteoarthritis (KOA), which is a degenerative joint disease of the knee that results from the progressive loss of cartilage and has a higher prevalence rate compared with other types of OA [2]. Obesity, age and previous injuries (e.g. ACL injury) due to sports or occupational/daily activities are factors that show a high correlation with KOA [3]. The quantification of KOA is performed with the Kellgren–Lawrence (KL) severity grading scale [4], which is the most commonly grading system (current gold standard) and consists of five severity grades, from 0 to 4 (Figure 2). At the onset of this disease, the main consequences are low quality of life due to pain, poor psychological state and social isolation. Due to KOA's multifactorial nature and the poor understanding of its

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

pathophysiology, there is a need for reliable tools that will reduce diagnostic and prediction errors made by clinicians. The existence of public databases (e.g. Osteoarthritis Initiative (OAI)) has facilitated the advent of advanced analytics in KOA research however the heterogeneity of the available data along with the observed high feature dimensionality make the prediction of KOA progression and the diagnosis tasks difficult [5, 6].

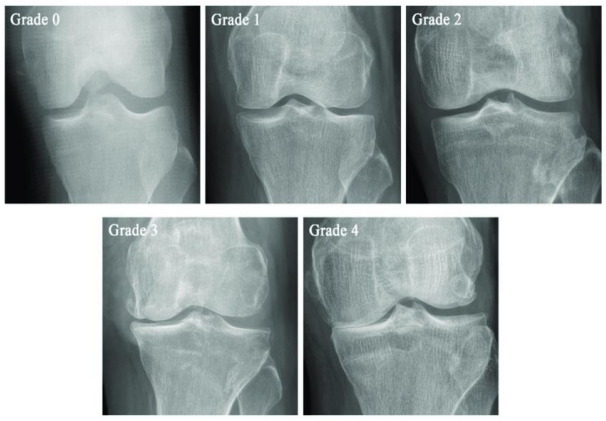


Figure 2. Kellgren-Lawrence (KL) scale [7].

As mentioned above, the existence of anterior cruciate ligament (ACL) injury is a risk factor which is high correlated with KOA. ACL tear is one of the most common knee injuries (Figure 3) and it results in knee instability and increased risk of early onset osteoarthritis. According to recent surveys, post-traumatic KOA has been observed in over 50% of individuals [8]. Specifically, 10 up to 20 years after anterior cruciate ligament reconstruction (ACLR) is the most common period for the occurrence of KOA. It is established that abnormal knee kinematics and kinetics after ACLR contribute to the degenerative processes, due to changes in cartilage loading. So, identifying significant gait changes is important for understanding normal and ACL function.

Consequently, the main challenges according to the existing literature are listed as follows:

- The existence of big data, requires advanced AI analytics tools,
- Big data show heterogeneity and high dimensionality, therefore robust feature selection techniques are required to cope with them,
- There is a need for prediction and diagnostic models that offer generalization to various data subsets and

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

- Despite the existence of prediction and diagnostic models, machine learning models function as black boxes, therefore there is a need to develop techniques for their interpretation.

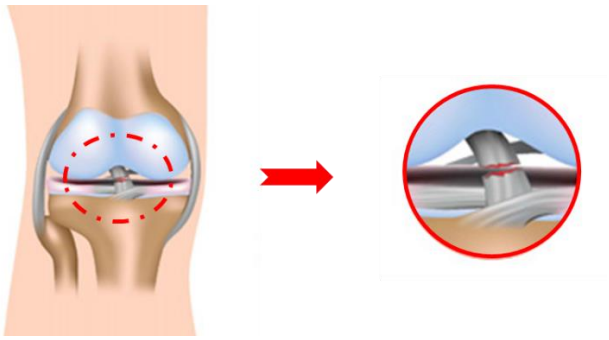


Figure 3. Anterior cruciate ligament (ACL) injury.

Meeting the aforementioned research challenges of KOA, the present PhD thesis incorporates five studies aiming i) to improve the current understanding of risk factors which have the main role in KOA progression and diagnosis tasks and to improve the current understanding of gait biomechanical parameters for ACL injury diagnosis, ii) to interpret their contribution on the model's output thus enhancing our understanding of the rationale behind the decision-making mechanism of the best model in each task and iii) to develop reliable and non-invasive tools for the prediction of KOA progression, KOA diagnosis as well as post-surgical rehabilitation tools.

As an introduction, in **Chapter 1**, a review is presented to introduce the reader to key directions of Machine Learning techniques on the diagnosis, predictions and post-treatment of KOA. As observed, KOA is a big data problem in terms of data complexity, heterogeneity and size. Hence, a gap was identified concerning the Machine Learning as the solution to cope with the aforementioned challenges and thus lead to new automated pre- or post-treatment solutions that utilize data from the greatest possible variety of sources. In **Chapter 2**, a robust feature selection (FS) approach that could identify important risk factors which contribute to the prediction of KOA with KL progression from a big pool of risk factors available in the osteoarthritis initiative (OAI) database was provided. Furthermore, three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression were explored and then machine learning-based models that can predict long-term KL progression were developed. In **Chapter 3**, an evolutionary genetic algorithm (GA)-based wrapper technique for the identification of risk factors for KOA progression was provided increasing the generalization capacity with respect to KOA progression. The proposed feature

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

selection (FS) methodology overcomes two crucial challenges: (i) the observed high dimensionality and heterogeneity of the available data that are obtained from the OAI database and (ii) the severe class imbalance problem posed by the fact that the KOA progressors class is significantly smaller than the non-progressors' class. According to the literature there is a need for reliable tools that will reduce diagnostic errors made by clinicians. The existence of public databases (i.e., OAI) has facilitated the advent of advanced analytics in KOA research however the heterogeneity of the available data along with the observed high feature dimensionality make this diagnosis task difficult. Hence in **Chapter 4**, a robust FS methodology based on fuzzy logic was provided. The proposed methodology has the aim: (i) to handle the multidimensional nature of the available datasets and (ii) to alleviate the defectiveness of existing feature selection techniques towards the identification of important risk factors which contribute to KOA diagnosis. In **Chapter 5**, an explainable ML-empowered methodology was provided to identify important biomechanical parameters associated with ACL injury diagnosis. In addition, a state-of-the-art explainability analysis based on SHAP and conventional statistical analysis attempted to uncover the rationale behind the decision-making mechanism of the best trained model and provide a holistic approach of quantifying the contribution of the input biomechanical parameters in the tasks of ACL injury diagnosis in order to avoid the outset of KOA in the future.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapter 1

Machine learning in knee osteoarthritis: A review

Published as:

Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. E. (2020). Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open, 2(3), 100069.

Abstract

The purpose of present review survey is to introduce the reader to key directions of Machine Learning techniques on the diagnosis and predictions of knee osteoarthritis. This survey was based on research articles published between 2006 and 2019. The articles were divided into four categories, namely (i) predictions/regression, (ii) classification, (iii) optimum post-treatment planning techniques and (iv) segmentation. The grouping was based on the application domain of each study. The survey findings are reported outlining the main characteristics of the proposed learning algorithms, the application domains, the data sources investigated and the quality of the results. Knee osteoarthritis is a big data problem in terms of data complexity, heterogeneity and size as it has been commonly considered in the literature. Machine Learning has attracted significant interest from the scientific community to cope with the aforementioned challenges and thus lead to new automated pre or post treatment solutions that utilize data from the greatest possible variety of sources.

Keywords: knee osteoarthritis; feature engineering; machine learning; prediction; classification; segmentation

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Introduction

Knee Osteoarthritis (KOA) is a degenerative disease of the knee joint and the most common form of arthritis causing pain, mobility limitation, affecting independence and quality of life in millions of people [2]. There is no known cure for KOA, but there are several medical, biological and environmental risk factors, both modifiable and non-modifiable, that are known to be involved in the development and progression of the disease [9]. The aforementioned data characterizing KOA are high-dimensional, heterogeneous and the limited number of simple logistic regression models are not capable of handling large numbers of risk factors and most importantly, any interactions between environmental and other medical and biological factors. Furthermore, they cannot identify the tendency of a healthy subject to show signs of the disease and its progression based on patient outcomes. Despite that, the power and importance of correct study design should not be underestimated. In the well-designed study even "simple" analysis can give trustful results. These significant shortfalls in OA risk prediction models require a completely different modelling and computational approach to the problem. Advanced machine learning techniques such as fuzzy-logic theory, discrimination metrics (e.g., mutual information gain indexes and Fisher discrimination ratios) and advanced classification models combined with novel and efficient feature selection methods suitable for very large data sets could significantly contribute to the problem of high dimensionality compared to the existing statistical techniques applied to the OA risk prediction problem.

Machine Learning (ML) is the study of how computer algorithms (i.e., machines) can “learn” complex relationships or patterns from empirical data and hence, produce (mathematical) models linking an even large number of covariates to some target variable of interest [10]. As mentioned before, the ability to analyze complex cases with a huge volume of data and the maximum possible results it renders ML a valuable tool against KOA. It is worth noting that ML has been applied in areas such as robotics [11], medicine [12], biochemistry [13], bioinformatics [14], meteorology [15], agriculture [16] and the economic sciences [17]. The importance of applying ML techniques to KOA has been documented by Jamshidi et al. [5] and Kluzek and Mattei [18] in 2019.

In this context this review has been carried out to allow each researcher to refer to the appropriate ML method in relation to KOA. To achieve this aim, the structure of the review is as follows. Section *Machine Learning in a nutshell* presents the terminology and definitions, the types, tasks and models, which are used in the studies on which this review was based. Section *Review of studies* presents the steps of the methodology that were followed for the collection and classification of the studies concerning ML techniques in KOA. In addition, it presents a summary of the studied literature,

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

highlighting the main characteristics of proposed ML approached divided into four categories. The review ends with Section *Discussion and Conclusions*, which mentions the future expectations and advantages that exist through the usage of machine learning in knee osteoarthritis.

Machine learning in a nutshell

In ML, a sample (e.g., a patient) is represented by a number of features which come in various forms and formats including patient's characteristics, risk factors, shape/texture characteristics in medical images or clinical history data. To facilitate the learning process, there features are typically concatenated forming a multidimensional feature vector. ML systems (Figure 1.1) operate in two phases: the learning phase (training) and testing one. Indicatively, the role of the pre-processing unit can be broadly categorised into the following: (i) data cleaning aiming to remove noise, missing and inconsistent examples (ii) data integration in cases where multiple data sources are available and (iii) data transformation including discretisation and normalisation. The feature extraction / selection unit (also referred as feature engineering unit) attempts to generate and/or identify the most informative feature subset in which the learning model will be subsequently applied during the training phase [19]. The feedback loop allows adjustments of the pre-processing and feature extraction / selection units that will further improve the performance of the learning model. During the testing phase, the trained model is shown previously unseen samples (represented as images or feature vectors) that need to be classified. The model makes an appropriate decision (classification or regression) based on the features that are present in each sample. Deep learning [20], that is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain, sets an alternative architecture by shifting the burden of feature engineering (the process of transforming raw data into features) to the underlying learning system. From this perspective, feature extraction or selection are omitted leading to a fully trainable system that begins from raw or pre-processed input (e.g., image pixels or time-series) and ends with the final output of recognized objects or predicted values.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

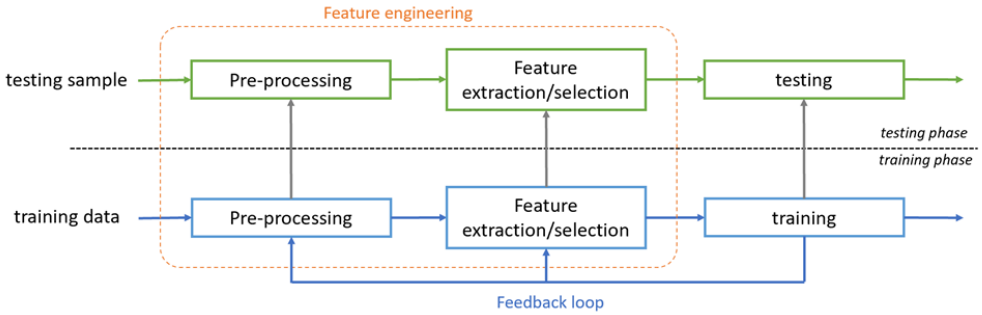


Figure 1.1. A typical machine learning system

Learning can be classified as supervised, unsupervised or reinforcement learning. In *supervised learning*, each data sample is represented by a pair consisting of an input (typically a multi-dimensional feature vector) and a desired output value (e.g., a label having real-world meaning such as Kellgren Lawrence grades in case of KOA). The training phase involves the task of learning a function that maps every input to its associated output. The generated inferred function is used to map unknown inputs during the testing phase. Unsupervised learning [21] is a class of ML techniques that operate with unlabeled data with the goal of discovering structures or patterns in the dataset. Novel paradigms for unsupervised learning (the so-called self-supervised learning) have been also proposed exploiting different labelings that are freely available besides or within visual data to learn general-purpose features [22]. In reinforcement learning, a model learns through trial-and-error interactions with its environment using reward and penalty assignments.

In the terminology of ML, classification is considered as an instance of supervised learning. In short, it is the task of identifying to which of a set of categories (sub-populations) a new example belongs, on the basis of a training set of data (experience) containing examples whose label is known. Regression constitutes another supervised learning task, which aims to provide a prediction of an output variable according to the input variables which are known. The most known regression algorithms are the linear regression [23], as well as, stepwise regression [24]. Also, more complex regression algorithms have been developed, such as ordinary least squares regression [25], multivariate adaptive regression splines [26], multiple linear regression, and locally estimated scatterplot smoothing [27]. Table 1.1 cites the most well-known state-of-the-art ML models of the literature. Dimensionality reduction (DR) is a task that belongs in both families of supervised and unsupervised learning types, with the aim of providing a more compact lower-dimensional representation of a dataset preserving as much information as possible from the original data. It is usually performed prior to applying a classification or regression model in order to avoid the effects of the curse of dimensionality. Some of the most common DR algorithms are

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

the following: (i) principal component analysis (PCA) [28], (ii) partial least squares (PLS) regression [29] and (iii) linear discriminant analysis (LDA) [30]. Finally, clustering [31] is an application of unsupervised learning typically used to find natural groupings of data (clusters). Well established clustering techniques are the K-means technique [32] hierarchical clustering [33], and the expectation-maximization technique [34].

Table 1.1. Presentation of indicative ML models along their characteristics.

Category	Models	Description	Advantages	Disadvantages
Bayesian	Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Bayesian Belief Network [35-37]	Probabilistic graphical models in which the analysis is undertaken within the context of Bayesian inference	They model uncertainty; easy to handle missing and hidden data	Increased computational cost in high-dimensional spaces; they require subjective definition of prior probabilities
Linear	Linear regression [23, 24]	The best fit line through all data points	Easy to understand and implement; models can be easily interpreted	Too simple to capture complex associations between variables: prone to overfitting
	Logistic regression [23]	The adaptation of linear regression in classification problems		
Tree-based [38-41]	Decision trees (DT) [42-44][37]	A decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility		Not powerful enough in problems of high complexity
	Random forest (RF) [41]	Ensemble model that produces multiple decision trees, using a randomly selected subset of training samples and variables.	Fast to train and powerful	Not so interpretable; slower than other techniques
	Gradient boosting [45]	Uses weak decision trees as base models. Predictive results are obtained through increasingly refined approximations.	Fast and high performing	Interpretability issues; sensitive to small changes
Neural networks	Neural networks [46-55]	Information processing paradigm that is	Can handle complex problems	Not interpretable; Slow
	Deep Neural networks (DNN) [20] such as CNN [56], deep belief network [57], and auto-encoders [58].	inspired by the way biological nervous systems, such as the brain, process information.	Can handle extremely complex problems	Require a lot of power; not interpretable; Slow
Instance based models	K-Nearest Neighbor [59], Locally Weighted Learning [60], Learning Vector Quantization	Memory-based techniques that learn by comparing new examples with instances in the training database	Simple and fast to implement	Complexity grows with data (up to O(n) where n is the number of the training examples), prone to overfitting

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Support vector machines (SVMs)	algorithm [61], Self-Organising Maps [62] SVM [63, 64] Least Squares SVM [65]	Finds a solution (linear or non-linear) that maximizes the margin between classes	SoA performance; generalized solutions; robust to high dimensionality	Tuning hyperparameters is crucial; time consuming and difficult to interpret

Recently, deep learning has attracted wide-spread attention because of its enormous representing power, automated feature learning capability and best-in-class performance in solving complex problems [66]. Deep NNs make use of deeper architectures, extensible hidden units and nonlinear activation functions to model complex data, whereas one of their most attractive aspects is that they automate feature engineering thus alleviating the need for domain expertise and hardcore feature extraction. Currently, DL models have dramatically improved the state-of-the-art in many different sectors and industries including healthcare [67]. DL models can be either supervised, partially supervised, or even unsupervised. Convolutional neural networks (CNN) are among the most famous DL networks where feature maps are extracted by performing convolutions in the image domain. A comprehensive introduction on CNNs is given in [56]. Other typical DL architectures that belong to the family of probabilistic undirected graphical models are deep Boltzmann machines, and deep belief networks [57]. Auto-encoders [58] are unsupervised DNNs whose main idea is to encode high dimensional data into a low-dimensional latent vector and try to reconstruct the input data as flawlessly as possible by only using its coding. Recurrent neural networks (RNN) are another important family of DL models that define unique topological connections between their neurons in order to encode temporal information in sequential data [68].

Methods

Literature Search Approach

This survey was based on research articles published between 2006 and 2020 using the search engines Scopus, PubMed and Google Scholar. During our search, we identified articles that used ML for the study of KOA by various techniques. Especially, for this search, the terms machine learning, deep learning and knee osteoarthritis were used. A prerequisite for the inclusion of an article in our research was the occurrence of one of the three terms mentioned as keywords, either in the title or in the abstract of each article.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Exclusion Criteria

In the first instance, all articles retrieved and collected were examined for the title and the abstract by one of the authors. In order to reach our original goal, we excluded the following categories: non-English articles, postgraduate dissertations, doctoral dissertations, studies not involving people with knee osteoarthritis and studies using traditional technical statistics. All the selected articles have been presented either in journal papers or conferences. Finally, the rest of the authors reviewed again the titles and abstracts to ensure that they met the membership criteria.

Assessed Outcomes

The studies, which are recorded in this article, were divided into four categories, namely (i) Predictions/Regression (13 studies), (ii) Classification (43 studies), (iii) Optimum post-treatment planning techniques (4 studies) and (iv) Segmentation (15 studies). The grouping was based on the technical characteristics of the ML methods and the application domain of each study.
Then, after separating the articles, the following information was extracted from each article: Author, Year of publication, Data (MRI, X-Ray, Kinetic and Kinematic data, Clinical data and Demographics), Feature Engineering approach, Learning Algorithm techniques, Validation and Results (evaluation of performance).

Results

Predictions/Regression

Despite the fact that OA field has been relatively slow adopting advanced analytical models compared to other fields, nowadays many studies focus on developing ML prediction models for KOA based on medical imaging (Magnetic Resonance Imaging (MRI), X-ray), clinical information, self-reported and biomechanical data.

Data sources: Imaging technologies (either MRI or X-ray) were incorporated into the majority of advanced analytical models to predict knee articular cartilage morphology with accuracies varying from 76.1% up to 92% [69-73]. Recently, the combination of multimodal data (medical images with clinical or biomechanical data) has formed the basis for more powerful and efficient models. To enhance the quality of the available raw data or overcome the curse of dimensionality, a number of sophisticated algorithms were reported in the literature including: (i) LASSO [74], Topological Data

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Analysis [75], Recursive feature elimination (RFE) [76], PCA [77] for dimensionality reduction or (ii) CNN [78] to extract new more informative deep features for images. The major finding of these studies was that the accuracy of image-based prediction of KOA progression can be improved if it is complemented with data sources such as clinical data, self-reported and biomechanical data.

Learning techniques: Due to their efficiency and predictive performance, ensemble algorithms (RF or Gradient Boosting) were selected in five out of the twelve (12) studies in this category. However, a significant number of studies employed simpler models (e.g., linear regression models [69, 79] or logistic regression [75]) to implement the regression or prediction task. Non-linear SVMs were also investigated in four (4) papers [71, 73, 76, 80] and this choice could be attributed to the fact that they are relatively efficient in low and medium size feature spaces and that they generalize well. More complex learning (and subsequently more difficult to handle) approaches were finally tested in some studies [71, 73, 74] using NN-based architectures such as Artificial neural networks (ANNs) and CNNs.

Validation: In the majority of those studies, validation has been performed with n-fold cross validation. Hold-out (typically 70%/30% for training/testing) and Leave-one-out cross-validation (LOOCV) have also been observed as a validation approach in some of the studies. It is worth noting that Tiulpin et al. [78] used an independent test set (acquired in another center) for validation. An overview with all the studies including prediction models of KOA are shown in Table 1.2:

Table 1.2. Studies with Predictions/Regression techniques.

Author	Year	Data	Feature engineering	Learning Algorithm	Validation	Results
Abedin, J. [74]	2019	Questionnaire data / X-ray	LASSO	Elastic Net (EN), Random Forests (RF) and a convolution neural network (CNN)	70% training/30% testing	Root Mean Square Error (RMSE) for the CNN, EN, and RF models was 0.77, 0.97 and 0.94 respectively
Ashinsky, B. G. [72]	2017	MRI	-	Weighted neighbor distance using compound hierarchy of algorithms representing morphology WN(D-CHRM)	LOOCV	75% acc

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Donoghue, C. [69]	2011	MRI	Laplacian Eigenmap Embedding	Multiple linear regression	270 knees as external validation group	Up to R^2= 0.75
Du, Y. [73]	2018	MRI	PCA	ANN, SVM, Random forest, Naïve Bayes	10-fold cross validation (10F-CV)	ANN with AUC= 0.761 for KL grade Random forest with area under the curve (AUC) = 0.785 for JSM
Du, Y. [71]	2017	MRI	PCA	ANN, SVM, Random forest, Naïve Bayes	10F-CV	receiver operating characteristic (ROC) AUC of 0.761 (ANN)
Halilaj, E. [79]	2018	X-rays and pain scores	-	LASSO regression	10F-CV for model selection and 10% for model evaluation	AUC of 0.86 for Radiographic progression
Lazzarini, N. [81]	2017	Clinical variables, food and pain questionnaires, biochemical markers (BM) and imaging-based information	Ranked Guided Iterative Feature Elimination, PCA	Random Forest	10F-CV	AUC of 0.823 by using only 5 variables
Marques, J. [70]	2013	MRI	Texture Analysis for extraction and Partial least squares (PLS) regression for selection	Fisher linear discriminant analysis	10F-CV for model selection. 10% for evaluation	ROC AUC of 0.92
Nelson, A.E. [77]	2019	Demographic, MRI and biochemical variables	Distance weighted discriminatio n (DWD), PCA	K- means, t-SNE	Validation on 597 participants-	z = 10.1 (z-scores)
Pedoia, V. [75]	2018	MRI and biomechanics multidimensional data	Topological Data Analysis	Logistic Regression	-	AUC 83.8%
Tiulpin, A. [78]	2019	X-ray, Clinical data	CNN	Logistic Regression (LR) and Gradient Boosting Machine (GBM)	OAI dataset for training and MOST dataset for testing, 5F-CV	AUC of 0.79
Widera, P. [76]	2019	Clinical and X-ray image assessment metrics	Recursive feature elimination	Logistic regression, KNN, SVC (linear kernel),	Standard 10-fold stratified cross-validation protocol	F1 score 0.573 - 0.689

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Yoo, T. K. [80]	2013	Kinematic data	-	SVC (RBF kernel) and RF SVM	Leave-one-out cross-validation (LOOCV)	97.4 % acc

Classification

This section presents the outcomes of our survey on the application of classification models on the field of KOA research. It is worthwhile to note the plurality of different datasets along with the heterogeneity of data types used by each study. The identified data sources are: biomechanical data (kinematic-kinetic data and EMG signals), osteoarthritic outcome score, demographic characteristics, some gene polymorphisms, radiographs, X-ray and MRI. For this reason, we are grouping the studies into two categories, the first for biomechanical data-scores and the second for images.

Biomechanical data and discrete variables

Data sources: Biomechanical data were the most widely used source of information in the reported studies including kinematic-kinetic data and electromyography signals. Furthermore, clinical data consisting of self-reported, osteoarthritic outcome scores, demographic characteristics and some gene polymorphisms were used as additional sources complementing the biomechanical features.

Feature engineering: Feature extraction and dimensionality reduction have been applied to improve the predictive capabilities of the learning models as well as to increase their computational efficiency. A variety of algorithms and techniques were reported in the literature including: (i) Simulated annealing (SA) [82], Genetic algorithms (GAs) [82], Discrete wavelet transform (DWT) [83, 84], Wavelet Packet [85], SVM-based Fuzzy criteria [85] and Mahalanobis Distance algorithm [86] for feature selection and/or extraction (ii) Probabilistic PCA (PPCA) [87] and PCA [88-91] for dimensionality reduction and (iii) feature subsets exploration or use of time-domain statistical features [92, 93] to lead in more powerful learning models. PCA has been observed to be the most popular feature engineering technique due to its simplicity and easiness to handle.

Learning techniques: A variety of machine learning models were used for implementing the detection and/or classification tasks. KNNs and SVMs were the most frequently selected algorithms being tested in (7) out of nineteen studies in this subcategory. Furthermore, RF [94], DT [86], Dempster Shafer Theory [82, 89], Bayes

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

classifier [87] and Discriminant analysis [88] were also investigated. Finally, the use of deep learning techniques (e.g., ANNs [89, 95], PNNs [96], MLPs [86, 93, 97] or CNNs [90, 93]) was limited due to the nature of the available training datasets (heterogeneous features and small sample sizes).

An overview of the aforementioned studies is shown in Table 1.3:

Table 1.3. Classification studies employing biomechanical data and/or distinct variables.

Author	Year	Data	Feature engineering	Learning Algorithm	Validation	Results
Aksehirli, Ö [96]	2013	Demographic characteristics and some gene polymorphisms	-	SVM, PNN	152 OA knees for training and 102 healthy for testing	76,77% acc & 90,55% acc
Beynon, M. J. [82]	2006	Biomechanical Data	Simulated annealing (SA) and genetic algorithms (GAs)	Dempster–Shafer theory of evidence (DST) & Linear discriminant analysis (LDA)	LOOCV	96.7% & 93.3% acc
de Dieu Uwisengeyimana, J. [93]	2017	Biomechanical Data	Time-domain statistical features	Multilayer perceptron, Quadratic support vector machine, complex tree & deep learning network with k-NN	22 subjects (11 healthy and 11 OA)	99.5%, 99.4% 98.3% & 91.3% acc
Deluzio, K.J. [88]	2007	Biomechanical Data	PCA	Discriminant analysis	CV	Misclassification rate 8%
Jones, L. [89]	2008	Biomechanical Data	PCA	The Dempster–Shafer (DS)-based classifier & ANN	LOOCV	97.62% & 77.82% acc
Kotti, M. [87]	2014	Biomechanical data	PPCA	Bayes classifier	47F-CV	82.62 % acc
Kotti, M. [94]	2017	Biomechanical data	-	Random forest	50% training/ 50% testing, 5F-CV	72.61% acc

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Lim J. [90]	2019	Demographic and personal characteristics, lifestyle- and health status-related variables	PCA	DNN	66% training/ 34% testing	AUC of 76.8%
Long, M. J. [98]	2017	Outcome scores and biomechanical gait parameters	-	KNN	70% training/30% test. 30% of training was left out for validation	AUC of 1.00
McBride, J. [99]	2011	Biomechanical data	-	Neural networks	50% training/50% testing	75.3% acc
Mezghani, N. [83]	2008	Biomechanical data	Discrete wavelet transform (DWT)	Nearest neighbor classification (NNC)	LOOCV	38 of 42 cases acc
Mezghani, N. [84]	2008	Biomechanical data	Discrete wavelet transform (DWT) & Polynomial expansion	Nearest neighbor classifier (NNC)	LOOCV	91% acc 67% acc
Mezghani N. [100]	2017	Biomechanical Data	-	Regression tree	10F-CV for model selection. 10% for model evaluation	ROC AUC of 0.85
Moustakidis, S. [85]	2010	Biomechanical data	Wavelet Packet, FS via SVMFuzCoC	KNN1 SVM (AAA) SVM (1AA) FCT C4.5 FDT-SVM	10F-CV	86.09 % acc 89.71 % acc 90.18 % acc 88.35 % acc 91.12 % acc 93.44 % acc

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Moustakidis, S. [92]	2019	Clinical Data	Feature subsets exploration	DNN Adaboost Fuzzy KNN Fuzzy NPC CFKNN	10F-CV	86.95% acc (for age 70+) 78.60% acc 77.39% acc 72.40% acc 73.60% acc
Phinyomark, A. [91]	2016	Biomechanical Data	PCA	SVM	10F-CV	98-100 % acc
Şen Köktaş, N. [97]	2006	Biomechanical data	-	MLPs	CV	1.5 of the subjects has been misclassified
Şen Köktaş, N. [86]	2010	Biomechanical data (Also included age, body mass index and pain level)	Mahalanobis Distance algorithm	Decision tree - MLP multi-classifier	10F-CV	80% acc
Yoo, T. K. [95]	2016	Predictors of the scoring system in the Fifth Korea National Health and Nutrition Examination Surveys (KNHANES V-1) data	Logistic regression	ANN	66.7% training /33.3% validation, KNHANES V-1 (internal validation group) and OAI (external validation group)	ROC AUC of 0.66-0.88

Medical Images

Medical images form a crucial source of information in the KOA research. The types of medical imaging that have been analysed in this survey were either MRI or X-ray.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

According to our knowledge only six studies have been presented in the literature, until now, that reported the development of MRI data analysis methodologies for the diagnosis of KOA. Only one of the aforementioned studies adopted a deep learning approach applying directly learning algorithms (CNN and specifically MRNet) on the available images without the inclusion of any feature selection technique [101]. The rest of the reported studies employed a number of feature engineering techniques prior to the application of the learning models. Discrete wavelet transform, Gray level Co-occurrence Matrix (GLCM) and PCA are among the algorithms that were used to either extract new features or reduce the feature space dimensionality. As regards the learning part, NNs [102, 103], SVM [104, 105] and LDA [106] were the most commonly employed models for early detection and diagnosis of KOA.

Localization of joints was a crucial task in the reported X-ray applications. Numerous approaches of varying complexity were applied such as filtering (Gabor, Sobel) [107-109], statistical shape/texture analysis [110, 111], fully automated software tools (Bonefinder [112, 113]) or more sophisticated deep learning networks including YOLO and FCN [114, 115]. In some cases, manual cropping was also performed [116-120]. PCA and GLCM were again selected in many of the reported papers to generate small and informative feature subsets, whereas several recent studies adopted CNN-based methodologies as an alternative for the feature extraction task. Deep learning networks (e.g., VGG-19, VGG-16, DenseNet, ResNet-34 and LSTM) were also involved in several studies acting as the main learning algorithm. State-of-the-Art ML models such as SVMs were finally selected in a few Xray-based studies to drive the decision-making process. In most of the cases, validation was performed via k-fold CV and hold-out whereas some studies adopted more robust validation strategies (cross-center validation). The main characteristics of the reported image-based classification studies are shown in Table 1.4.

Table 1.4. Medical image-based classification studies of KOA.

Author	Year	Data	Localization of joints	Feature engineering	Learning Algorithm	Validation	Results
Bien, N. [101]	2018	MRI	-	-	CNN (MRNet)	Validation A: 82,9% training, 8.5% tuning and 8,6 validation Validation B: 60%-20%-20% into training, tuning, and validation	AUC of 0.937

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

						sets using an external dataset	
En, Chuah Zhi [102]	2013	MRI	-	Discrete Wavelet Transform (DWT)	ANN-based	57,1% (200 images) training/ 42,9% testing (150 images)	94.67% acc
Kubkaddi, Sanjeevakumar [104]	2017	MRI	-	GLCM	SVM with RBF kernel, SVM with linear kernel & SVM with polynomial kernel	70% training/ 30% testing	95.45% acc, 95.45% acc & 87.8% acc
Kumarv, A. [105]	2017	MRI	-	GLCM	SVM	15 images / hold out validation	86.66% acc
Marques, J. [106]	2012	MRI	-	PLS with forward feature selection (PLS-FFS)	Fisher LDA, PLS regression, sparse PLS and sparse LDA	10F CV	ROC AUC of 0.86 (Diagnosis) & 0.63 (Prognosis), ROC AUC of 0.88 & 0.67, ROC AUC of 0.89 & 0.69, ROC AUC of 0.93 & 0.70, ROC AUC of 0.89 & 0.59
Pedoia, V. [103]	2019	MRI (T_2 relaxation time maps), Demographics and KOOS	-	PCA	Densely Connected Convolutional Neural Network (DenseNet) RF	65-20-15% split of training, validation, and holdout testing set	AUC = 83.44%, Sensitivity = 76.99%, Specificity= 77.94% AUC = 77.77%, Sensitivity = 67.01%, Specificity = 71.79%

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Anifah, L. [107]	2013	X-ray	Gabor filter	GLCM	Self Organising Maps (SOM)	16,2% training/ 83,8% testing	Accuracy rate of 93.8% for KL-Grade 0, 70% for KL-Grade 1, 4% for KL-Grade 2, 10% for KL-Grade 3 and 88.9% for KL-Grade 4
Anifah, L. [109]	2018	X-ray	Gabor kernel	-	SOM	8,8% training/ 91,2% testing	40.52% acc for KL-Grade 2 & 36.21% for KL-Grade 0
Antony, J. [121]	2017	X-ray	FCN	FCN	CNN	70% training/ 30% validation, Multi-center validation	Multi-class classification accuracy 60.3%
Antony, J. [108]	2016	X-ray	Sobel horizontal image gradients, linear SVM	Pre-trained CNN (BVLC reference CaffeNet and VGG-M-128 networks)	Linear SVM	70% training (with 5F CV)/ 30% testing	Fitting a linear SVM produced 95.2% 5F CV and 94.2% test accuracy for knee joint detection; 57.6% accuracy in the muti-class KOA severity task (Grades 0-4)
Bayramoglou, N. [112]	2019	X-ray	BoneFinder	Local Binary Patterns (LBP), Fractal Dimension (FD), Haralick features, Shannon	Logistic regression	5F CV on OAI for training and validation in MOST data	AUC of 0.84

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

				entropy, and Histogram of Oriented Gradients (HOG)			
Chen, P. [114]	2019	X-ray	Customized one-stage YOLOv2 network	-	CNN models (VGG-19)	training, validation, and testing sets with a ratio of 7: 1 :2.	69.7% acc
Gorriz, M. [122]	2019	X-ray	Trainable attention modules		CNN (VGG-16)	70% training/30 % testing and 10% of the training data was kept for validation	64.3% acc
Gornale, Shivanand S. [116]	2017	X-ray	Images are cropped to 512x409 pixels and finally rescaled	Histogram of orientated gradients (HOG)	Multiclass SVM	Classification results validated by two experts that were in close agreement	Classification rate of 97.96% for Grade-0, 92.85% for Grade-1, 86.20% for Grade-2, 100% for Grade-3 & Grade-4
Liu, B. [123]	2020	X-ray	Region proposal network (RPN)	-	FLA (Faster R-CNN as original and our adjusted model as FLA)	5F CV	82.5% acc
Minciullo, L. [110]	2017	X-ray	PCA based combination of statistical shape and texture models	PCA-3 stage Constrained Local Model	Indecisive Forest (IF) Optimised Indecisive Forest (OIF)	5F CV	87.61 % acc 88.15 % acc
Minciullo, L. [111]	2017	X-ray	Shape Model	Statistical Shape Model (PCA)	Random Forest	5F CV	ROC AUC of 0.842 (binary) & 0.479 (5-class problem)
Navale, D. I. [117]	2016	X-ray	Dividing Image into Blocks	Texture analysis algorithm	SVM	71,4% training, 4,8% validation	For affected subjects' accuracy is 80%

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

						and 23,8% testing	
Sharma, S. [118]	2016	X-ray	Cropping of images	Histogram method, GLCM and Canny Edge Detection Technique	SVM	75% training/ 25% testing.	95% acc
Tiulpin, A. [115]	2018	X-ray	FCN, as proposed in Antony 2017	-	CNN ResNet-34	67% training, 11% validation and 22% testing, multi-center validation	66.71% acc (multi-class Grades 0-4)
Tiulpin, A. [113]	2019	X-ray	Random forest regression voting approach implemented in a BoneFinder tool	-	An ensemble of deep residual networks with 50 layers, squeeze-excitation and ResNeXt blocks	5-fold subject-wise stratified CV	AUC of 0.98
von Tycowicz, C. [124]	2019	X-ray	-	Shape Space, Graph Convolutional Filters	A multi-layer, feed-forward graph convolutional network	The data was split into training, validation, and test sets with a ratio of 2/3, 1/6, and 1/6, respectively	64.64% acc
Wahyuningrum, R. T. [119]	2016	X-ray	Images were cropped around the knee properly	Contrast Limited Adaptive Histogram Equalization (CLAHE)-2DPCA/ Structural 2-Dimensional Principal Component Analysis (S2DPCA)	SVM (Gaussian kernel)	3F CV	Up to 94.33% class accuracy for Grade 0
Wahyuningrum, R. T. [120]	2019	X-ray	Manually cropping on the knee	CNN (VGG-16)	Long Short Term	3F CV	75.28% acc

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

	joint with dimensions of 400 x 100 pixels	Memory (LSTM)

Optimum post-treatment planning techniques

As concluded in this survey, there is a lack of studies on the development of ML based decision support systems (DSS) for the post-treatment stage of KOA. According to our knowledge, the first attempt in that direction was made in 2009 in [125] where the authors presented an approach for detecting recovery from knee replacement surgery using gait spatio-temporal parameters. Their main aim was to investigate if the classifier could detect changes at 2 and 12 months following knee replacement surgery. The proposed method achieved to: (i) detect improvements in gait function and (ii) recognize gait parameters that are altered due to KOA. In [126], the authors tackled the task of selecting the appropriate gait re-training strategy as a ML problem and presented interpretable learning models. Using the trained models, a specialist was able to know which technique would work best for a specific patient. Online segmentation for KOA rehabilitation monitoring was also investigtaed in [127]. The novelty of this system was the real-time feedback to patients and physiotherapists. Finally, an SVM-based human motion identification for rehabilitation exercise assessment of KOA was proposed in [128] using biomechanical data with reliable results (up to 100% in recognizing the types of rehabilitation exercises and over 97.7% in motion identification). In the majority of the reported studies, the SVM technique was applied (in three out of four reports) on biomechanical data leading to even perfect identification rates (up to 100%). The validation was performed with 10-fold cross validation or with the leave one out (LOO) cross-validation approach. The studies with the ML-empowered post-treatment planning techniques of KOA are shown in Table 1.5.

Table 1.5. Studies with ML-driven post-treatment planning techniques of KOA.

Author	Year	Data	Feature engineering	Learning Algorithm	Validation	Results
Chen, H. P. [127]	2016	Biomechanical data	Tilt angle calculation and initial posture classification algorithm	Multi-layer SVM	10-fold cross validation	90.6% on layer-1 SVM & 92.7% on layer-2 SVM

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Huang, P. C. [128]	2017	Biomechanical data	Sequential forward feature selection (SFS)	Multi-class SVM	10-fold cross validation	Accuracy for rehabilitation exercises recognition is 100% and for motion identification is 97.7%.
Levinger, P. [125]	2009	Biomechanical data	SVM	SVM	LOOCV	Accuracy of 100% for the training set and 88.89% for the test set
Wittevrongel, B. [126]	2015	Biomechanical data	k-equal frequency binning	Decision tree & Rule sets	LOOCV	Best accuracy 92.9% & 76.5 % respectively

Segmentation

Image segmentation is the process of changing the representation of an image into meaningful segments. MR image segmentation for KOA is typically performed by clinicians following a manual, laborious, time-consuming process that is prone to subjective diagnosis error. Therefore, many studies have focused on interactive, semi or fully automated cartilage segmentation to assist the medical research in KOA. At this point, it should be mentioned that even in the case of ML and especially in supervised learning approaches, a researcher/doctor still needs to label the images, hence the developed trained model is prone to the subjectivity.

Landmark localization and shape modelling: To increase the performance of medical image segmentation techniques, landmark localization and shape modelling have been utilized as preliminary tools before the application of ML or DL. As recorded, landmark localization took place by using either hourglass-like encoder-decoder models or with manual cropping and selection of seed points. Furthermore, a number of shape modelling tools were employed to extract informative shape-relevant characteristics from the available images including: Statistical Shape Models (SSMs), Combined Intensity, Shape Priors, Histogram of Oriented Gradients (HoG) and edge detectors.

Segmentation: Segmentation was accomplished employing either interactive or (semi- and/or fully) automated approaches. Flexible seeds labelling applied on MRI data [129] was the dominant approach on the integrative segmentation category. To enable automation on the segmentation tasks, advanced DL-based techniques were adopted (e.g., CNN [130-132], unsupervised domain adaptation DL [133] and DNN [134] or even state-of-the-art ML techniques such as SVM [135], KNN [136, 137] and RF [138, 139]). Finally, more traditional segmentation approached were also proposed including: two-pass block discovery mechanism [140], Iterative Local Branch-and-mincut [141], Gaussian fit model [142] and multi-atlas segmentation (MAS) [143].

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Validation: OAI and MOST were the most-used databases to validate the performance of the aforementioned segmentation approaches. Validation was performed using k-Fold CV, LOOV or even manual assessment from experts.

An overview of all the identified KOA segmentation studies of our survey is given in Table 1.6:

Table 1.6. Segmentation techniques applied on the KOA research.

Author	Year	Data	Feature engineering	Learning Algorithm	Validation	Results
Ababneh, S. Y. [140]	2010	MRI	Disjoint (non-overlapping) block-wise scanning, two-pass block discovery	A graph-cut based segmentation algorithm	30 images from the OAI database	96 % acc
Ambellan, F. [130]	2019	MRI	-	Combination of Statistical Shape models (SSMs) and 2D / 3D CNN	Datasets: (i) SKI10, (ii) OAI Imorphics and (iii) OAI ZIB	(i) 74.0 ±7.7 Total score (ii) For femoral cartilage the DSC is 89.4%; for baseline and 89.1% (iii) The DSC is 98.6% for femoral bone, 98.5% for tibial bone, 89.9% for femoral cartilage, and 85.6% for tibial cartilage
Gan, H. S. [129]	2017	MRI	k-means clustering algorithm, Fuzzy c-mean	Flexible seeds labelling method	Manual validation by two experts on 10 images	Dice's reproducibility of 0.80 for observer 1 and 0.82 for observer 2
Gornale, Shivanand S. [136]	2019	X-Ray	ROI extraction using Sobel, Prewitt edge detection, Computation of basic statistical features	Otsu's Segmentation, Texture based Segmentation and KNN	532 digital Knee X-ray images	The accuracy rate of 91.16% for Sobel method, 96.80% for Otsu's method, 94.92% for texture method and 97.55% for Prewitt method is obtained
Kashyap, S. [138]	2016	MRI	Extraction of 3D Haar-like features from volume of interest (VOI)	LOGISMOS, just-enough interaction (JEI) as post-processing and Random Forest Classifier	The data from OAI were divided into two training sets with 15 and 13 which were used to train the NAF and the second RF classifier. 53 data-sets were used for testing	Border positioning errors (mm) Femur signed 0.03±0.19 Femur unsigned 0.55±0.11 Tibia signed 0.10±0.17 Tibia unsigned 0.61±0.14, For RF classifier: Femur signed -0.06±0.18 Femur unsigned 0.56±0.11 Tibia signed 0.16±0.24 Tibia unsigned 0.65±0.17
Kashyap, S. [139]	2018	MRI	Neighborhood Approximation Forests k-means clustering	Hierarchical Random Forest Classifier and LOGISMOS	108 MRIs from baseline, and 12-month follow-up scans of 54 patients	Cartilage surface positioning errors (in mm) of 4D Femur signed 0.01±0.18

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

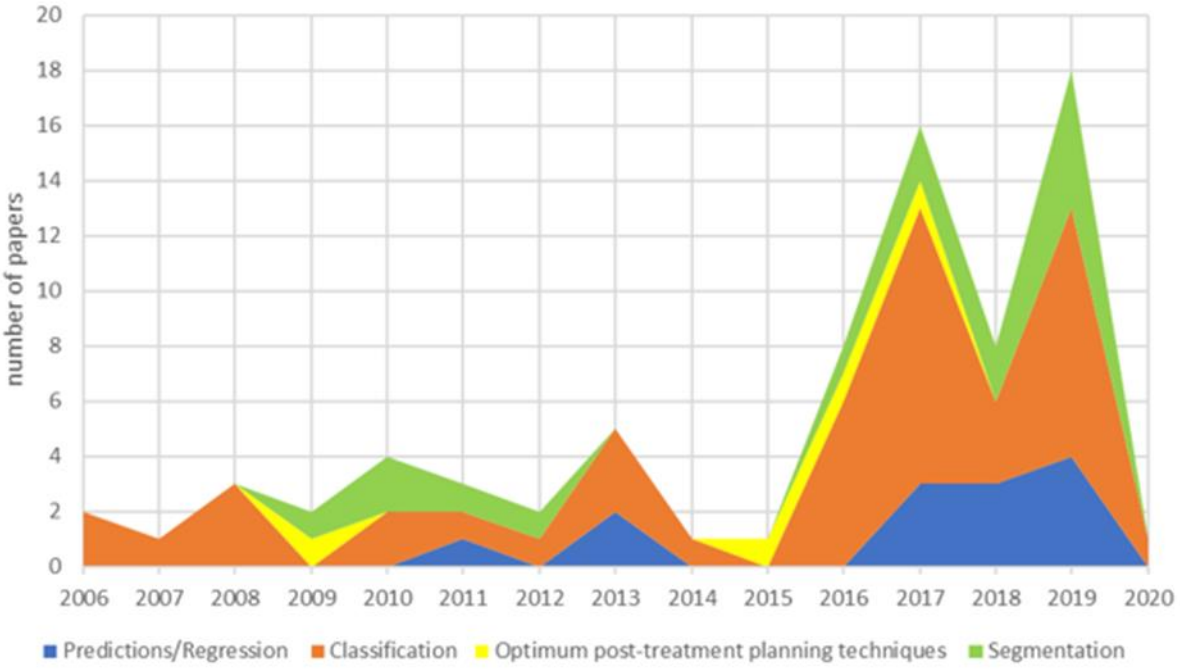
Marstal, K. [137]	2011	MRI	Histogram equalization, extraction of similarity features from neighboring patches and PCA	K-means	MRI scans from 50 subjects (25 for training)	Femur unsigned 0.53±0.11 at Baseline Average sensitivity, specificity and dice similarity coefficient of 0.853 ± 0.093, 0.999 ± 0.001, 0.800 ± 0.106 and 0.831 ± 0.095, 0.999 ± 0.001, 0.777 ± 0.054 on tibial and femoral cartilages respectively
Panfilov, E. [133]	2019	MRI	-	Deep learning U-net with two modern regularization techniques, namely, supervised mixup and UDA	5-fold cross-validation. Dataset A: 88 MRI images, Dataset B: 108 MRI images and Dataset C: 44 MRI images	Mean of volumetric DSCs is 0.907 (U-net + mixup, Dataset A) for femoral cartilage and DSCs is 0.821 (U-net + UDA2, Dataset C).
Park, S. H. [141]	2009	MRI	Combined Intensity and Shape Priors	Iterative Local Branch-and-mincut	LOOV on 8 3D MRI images	Average similarity index over 0.80 for normal participants and 0.75, 0.67, and 0.64 for participants with established knee OA
Swanson, M. S. [142]	2010	MRI	Manual selection of seed points, histogram and fitted Gaussian curves of the region	Threshold operation followed by conditional dilation and post-processing	Validation on 10 normal knees images and 14 knees with OA	Mean similarity Index 0.64-0.80
Tack, A. [131]	2018	MRI	2D U-net followed by statistical shape models of menisci	CNN (3D U-Net)	Validation on 5 different datasets of MRI images from OAI with 2F CV	DSCs was 83.8% for medial menisci (MM) and 88.9% for lateral menisci (LM) at baseline, and 83.1% and 88.3% at 12-month follow-up.
Tack, A. [132]	2019	MRI	-	3D CNN (3D U-Nets)	MRI data of 1378 subjects from the OAI (2F CV)	Accuracy of 88.02 ± 4.62 for medial tibial cartilage (MTC) and 91.27 ± 2.33 for lateral tibial cartilage (LTC) at baseline and 87.43 ± 4.02 and 90.78 ± 2.42 at 12-months follow-up
Tamez-Pena, J. G. [143]	2012	MRI	Manual creation of atlases by experts using CiPAS	Multi-atlas segmentation using CiPAS platform	LOO on 48 MRI images	DSC 0.88 and 0.84 for the femoral and tibial cartilage
Tiulpin, A. [135]	2017	X-ray	Anatomically-based joint area proposal and Histogram of Oriented Gradients	SVM	The images from MOST were used to create training (991), validation (110) and test sets (473), Jyvaskyla (93), OKOA (77)	Mean intersection over the union equals to: 0.84 (MOST), 0.79 (Jyvaskyla) and 0.78 (OKOA).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Tiulpin, A. [134]	2019	X-ray	ROI localisation using low-costs annotations	Hourglass-like encoder-decoder models for landmark localization	5-fold patient-wise cross-validation split stratified by a KL grade (748 knee joints in total)	Presicion 92.11 ± 0.34 at 2.5mm

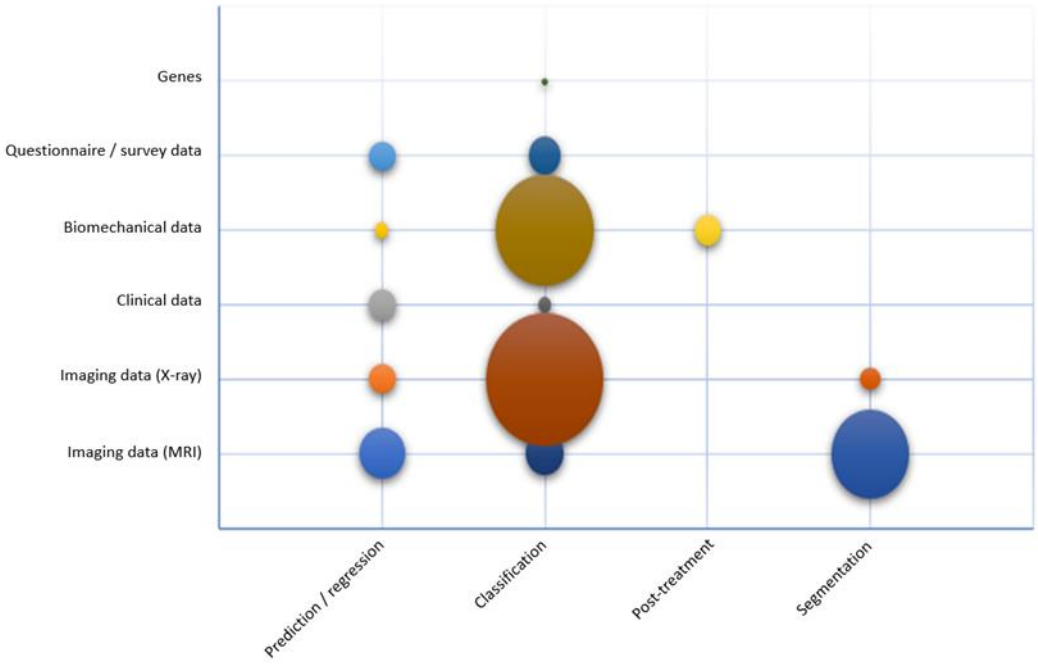
Discussion and Conclusions

Our literature survey outlined the current usage of machine learning methods in KOA diagnosis and prediction challenge. Figures 1.2 shows an increasing trend of ML-related studies and papers in the field of KOA indicating the need for (i) enhancing our understanding about the onset and progression of the disease and (ii) new data-driven tools that could enable early diagnosis and prediction of KOA. ML could play a key role towards these directions extracting valuable knowledge from various types of clinical data (biomechanical parameters, images, kinematics) and finding new solutions that utilize data from the greatest possible variety of sources.



(a)

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186



(b)

Figure 1.2. a) A temporal evolution chart depicting the number of papers per category published each year since the year 2006 and included in the survey, b) Bubble chart showing a distribution of the papers considered in this survey arranged according to the data sources utilized in each survey category.

Data has to be seen as an asset being one of the most important and instructive assets of the healthcare industry. In KOA research, several data sources have been considered as inputs forming powerful multi-dimensional training and testing data sets. Medical Imaging is one of the dominant data sources of the sector with MRI and X-ray images being typically employed in the majority of the papers of our survey (25 and 25 papers out of 75 used MRI and X-ray, respectively). Biomechanical parameters were also investigated in 21 studies demonstrating a big potential to be useful input data in KOA diagnosis, prognosis and the post-treatment planning. Finally, other complementary data sources have been also considered in KOA research in several papers including pain, outcome scores, demographics, generic attributes and genes (Figure 1.2).

Feature engineering algorithms were applied on the available clinical data to either reduce the input feature dimensionality or extract new informative parameters from the raw data. PCA was employed in a number of papers to compress 3D kinematic time-series, ground reaction forces and MRI/X-ray images into more compact representations. Time domain and time-frequency domain features (e.g., DWT or Wavelet packet) were also extracted from GRF or EMG signals. GLCM was proved to be a quite popular technique for extracting textural features in studies where MRI or

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

X-ray images are considered as inputs. A number of feature selection techniques has been also employed to select the most informative features from the pool of the available or extracted parameters. Partial least squares, simulated annealing, random selection and sequential forward FS were among the techniques that were used to reduce the feature dimensionality of the initial space so as to increase the computational efficiency as well as generalisation capability of the subsequent classification or regressing models. Pre-trained CNN models were finally employed to extract valuable information for clinical images.

As far as the type of the ML models that were reported in our survey, SVMs were proved to be the most frequently used model in all the survey categories. Four (4) SVM-based studies were identified in the knee OA prediction survey, whereas another ten (10) papers made use of SVM for classification purposes including biomechanical discrete parameters or images (mostly MRI and X-ray). Moreover, SVM was also employed in three (3) out of the four (4) papers in the post-treatment survey. The choice of SVM could be attributed to the fact that they generalize well in practice and that are computationally effective in high dimensional spaces. Neural networks were the second most frequent technique with three (3) studies reported for knee OA prediction and eighteen (18) applications of NN-based models in the OA classification survey. Convolutional neural networks were finally considered in studies where clinical images were used as inputs. CNN-based approaches were either employed for feature extraction and/or for quantifying the severity of knee OA.

Nowadays biomedical research and clinical practices on KOA are struggling to cope with the growing complexity of interactions with the gained knowledge being fragmented and associated either with molecular/cellular processes or with tissue and organ phenotype changes related to clinical symptoms. Therefore, KOA is a big data problem in terms of the big data complexity and not the data size as it has been commonly considered in the literature. To tackle this huge complexity challenge, a multidisciplinary research approach should be proposed in the future across many disciplines: biomedical modelling via mechanistic analyses at various scales to capture locally the available knowledge into predictive simulations; medical imaging and sensing technologies to produce quantitative data about the patient's anatomy and physiology: data processing to extract from such data information that in some cases is not immediately available; big data analytics and computational intelligence tools that will generate personalised 'hyper-models' under the operational conditions imposed by clinical usage. Machine learning can explore massive design spaces to identify correlations and multiscale modelling can predict system dynamics to identify causality. This has the potential to lead to the development of individually tailored treatments to maximize the efficacy of treatment. Research work at the intersection of machine learning and KOA offers great promise for improving clinical decision-

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

making, and accelerating relevant intervention programs. To enable appropriate adoption of advanced learning algorithms and stay tuned with the new developments in ML/DL that are embracing research to other medical fields, open data, tools, and discussions must be forceful encouraged within the KOA research community.

Declaration of Competing Interest

None.

Acknowledgments

This work has received funding from the European Community's H2020 Programme, under grant agreement Nr. 777159 (OACTIVE).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapter 2

Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients

Published as:

Kokkotis, C., Moustakidis, S., Giakas, G., & Tsaopoulos, D. (2020). Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients. Applied Sciences, 10(19), 6797.

Abstract

Knee Osteoarthritis (KOA) is a multifactorial disease that causes low quality of life, poor psychology and resignation from life. Furthermore, KOA is a big data problem in terms of data complexity, heterogeneity and size as it has been commonly considered in the literature with most of the reported studies being limited in the amount of information, they can adequately process. The aim of this work is: (i) To provide a robust feature selection (FS) approach that could identify important risk factors which contribute to the prediction of KOA and (ii) to develop machine learning (ML) prediction models for KOA. The current study considers multidisciplinary data from the osteoarthritis initiative (OAI) database, the available features of which come from heterogeneous sources such as questionnaire data, physical activity indexes, self-reported data about joint symptoms, disability and function as well as general health and physical exams' data. The novelty of the proposed FS methodology lies on the combination of different well-known approaches including filter, wrapper and embedded techniques, whereas feature ranking is decided on the basis of a majority vote scheme to avoid bias. The validation of the selected factors was performed in data subgroups employing seven well-known classifiers in five different approaches. A 74.07% classification accuracy was achieved by SVM on the group of the first fifty-five selected risk factors. The effectiveness of the proposed approach was evaluated in a comparative analysis with respect to classification errors and confusion matrices to confirm its clinical relevance. The results are the basis for the development of reliable tools for the prediction of KOA progression.

Keywords: knee osteoarthritis; prediction; feature selection; machine learning; clinical data; KL-grade

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Introduction

Knee Osteoarthritis (KOA) is the most common type compared with other types of osteoarthritis (OA). KOA results from a complex interplay of constitutional and mechanical factors, including mechanical forces, local inflammation, joint integrity, biochemical processes and genetic predisposition. The specific disease causes significant problems when it occurs. In recent years, it has been also realized that KOA is closely associated with obesity and age [3]. Moreover, KOA is diagnosed in the young and athletes following older injuries [144]. The particularity of this disease is that the knee osteoarthritic process is gradual with a variation in symptoms intensity, frequency and pattern [2]. Due to the multifactorial nature of KOA, disease pathophysiology is still poorly understood and prognosis prediction tools are under current investigation.

Prognosis and treatment of KOA is a challenge for the scientific community. Increasing data collection has led to an increasing number of studies employing big data and AI analytics applied in the KOA research. As a result of this, several techniques have been reported in the literature in which ML models were used to predict KOA [6]. In 2017, Lazzarini et al. developed five (5) ML models that can be used to predict the incidence of knee OA in overweight and obese women. By integrating a wide variety of biomedical data in their models, they showed that using a small subset of the available information is possible to accurately predict the incidence of KOA by using Random Forest (RF) [81]. In another study, Halilaj et al. aimed to characterize different clusters of KOA progression and build models to predict these clusters early [79]. LASSO regression models were used to predict joint space narrowing and pain progression which are the most widely used surrogates of structural and symptomatic disease status. Furthermore, Pedoia et al. [75] used MRI and multidimensional biomechanics data attempting to meet the existing gap in multidimensional data analysis for precision medicine in KOA. They achieved large-scale integration of compositional imaging and skeletal biomechanics by using logistic regression as the ML model.

In 2019, Abedin et al. built two different prediction models, which achieved comparable accuracy with the aforementioned studies. In this study elastic net and RF were used along with a convolution neural network. The aim of this work was to explore whether the prediction accuracy of a statistical model based on the patient's questionnaire data is comparable to the prediction accuracy based on X-ray image-based modeling to predict KOA severity [74]. In another study, in 2019 Nelson et al. applied innovative ML approaches (e.g., K- means, t-SNE), specialized for a high dimension, low sample size setting, to phenotyping in KOA in order to better define progression phenotypes that may be more homogeneous and responsive to potential disease modifying interventions [77]. Moreover, in 2019 Tiulpin et al. proposed a novel

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

method based on ML that directly utilizes raw radiographic data, physical examination, patient's medical history, anthropometric data and, optionally, a radiologist's statement (Kellgren and Lawrence (KL)-grade) to predict structural KOA progression by using logistic regression and gradient boosting machine. They demonstrated that a knee X-ray image alone is already a very powerful source of data to predict whether a particular knee will have OA progression or not [78]. Futhermore, in the same year, Widera et al. used several ML models (e.g., logistic regression, K-nearest neighbor, SVC (linear kernel), SVC (RBF kernel) and RF) in combination with clinical data and X-ray image assessment metrics to develop predictive models for patient selection that outperform the conventional inclusion criteria used in clinical trials [76]. However, few studies have tried to apply ML models for the prediction of KOA. There is still a lack of knowledge on the contribution of self-reported clinical data on the KOA prognosis and their impact on the training of the associated ML predictive models [69, 70, 72, 80, 92, 145].

According to our knowledge, identification of risk factors for developing and especially predicting KOA has been limited by an absence of non-invasive methods to inform clinical decision making and enable early detection of people who are most likely to progress to severe KOA. Hence the main purpose of this study is twofold: (i) The prediction of KOA through the identification of risk factors that are relevant with KL progression from a big pool of risk factors available in the osteoarthritis initiative (OAI) database and (ii) the development of machine learning-based models that can predict long-term KL progression. To accomplish the aforementioned targets, a robust ML pipeline that involves a hybrid feature selection technique and well-known ML models was implemented. Moreover, this work also explores three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. Finally, a discussion on the nature of the selected features is also provided.

Data Description

Data were obtained from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/). Specifically, the current study only includes clinical data from: (i) The baseline; (ii) the first follow up visit at month 12 and (iii) the next follow up visit at month 24 from all individuals being at high risk to develop KOA or without KOA. Eight feature categories were considered as possible risk factors for the prediction of KL as shown in Table 2.1. Furthermore, our study was based on Kellgren and Lawrence (KL) grade as the main indicator for assessing the clinical status of the participants. Specifically, the variables 'V99ERXIOA' and 'V99ELXIOA'

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

were used to assign participants into subgroups (classes) of participants whose KOA status progresses or not (during labelling process).

Table 2.1. Main categories of the feature subsets considered in this work.

		Timeline of Visit		
Category	**Description**	**Baseline**	**12 Months**	**24 Months**
Subject characteristics	Anthropometric parameters including height, weight, BMI, abdominal circumference, etc.	●	●	●
Behavioural	Participants' social behaviour and quality level of daily routine	●	●	●
Medical history	Questionnaire data regarding a Participant's arthritis-related and general health histories and medications	●	-	-
Medical imaging outcome	Medical imaging outcomes (e.g., osteophytes and joint space narrowing)	●	-	-
Nutrition	Block Food Frequency questionnaire	●	-	-
Physical activity	Questionnaire results regarding leisure activities, etc.	●	●	●
Physical exam	Physical measurements of participants, including isometric strength, knee and hand exams, walking tests and other performance measures	●	●	●
Symptoms	Arthritis symptoms and general arthritis or health-related function and disability	●	●	●

In this work, we consider KL grades prediction as a two-class classification problem. Specifically, the participants of the study were divided into two groups: (1) Non-progressors: Healthy participants (KL grade 0 or 1) that remained healthy throughout the whole duration of the OAI study (eight years) and (2) KOA progressors: Healthy participants who developed OA (KL > 1) during the curse of the OAI study. So, the main objective of the study is to build ML models that could discriminate the two aforementioned groups and therefore be able to decide whether a new testing sample (healthy participant) will develop OA (assigned in the progressors' class) or not (assigned to the non-progressors' class). Secondary objectives of the study are to: (i)

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Identify which of the available risk factors contribute more to the classification output and as result can be considered as contributing factors in the prediction of OA and (ii) explore three different options (a single visit, two visits within a year and two visits within two years) with respect to the time period within which data should be considered in order to reliably predict KOA progression. To achieve these targets, we have worked on five different approaches in which different data subsets were considered comprising features from the baseline combined (or not) with features from visits 1 (at month 12) and 2 (month 24). The motivation behind this is to investigate whether data from the baseline are sufficient to predict the progression of KOA or additional data from subsequent visits should be also included in the training to increase the predictive accuracy of the proposed techniques. Detailed information as far as the aforementioned data subsets is given in the following. Data resampling was applied at each of the five datasets to cope with the problem of class size imbalance and generate dataset in which classes are represented by an equal number of samples.

- Dataset A (FS1): Progressors vs non-progressors using data from the baseline visit

Input: This dataset only contains data from the baseline (724 features). After data resampling, the participants were divided into two equal categories (Figure 2.1), as follows:

- Class A1 (KOA progressors): This class comprises 341 participants who had KL 0 or 1 at baseline, but they had also some incident of KL ≥ 2 at visit 1 (12 months) or later until the end of the OAI study in at least one of the two knees or in both.
- Class A2 (non-progressors): This class involves 341 participants with KL 0 or 1 at baseline, with follow-up x-rays but no incident of KL ≥ 2 for both of their knees until the end of the OAI study.

Output: Classification outputs 0 and 1 corresponding to assignments to classes A1 and A2, respectively.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

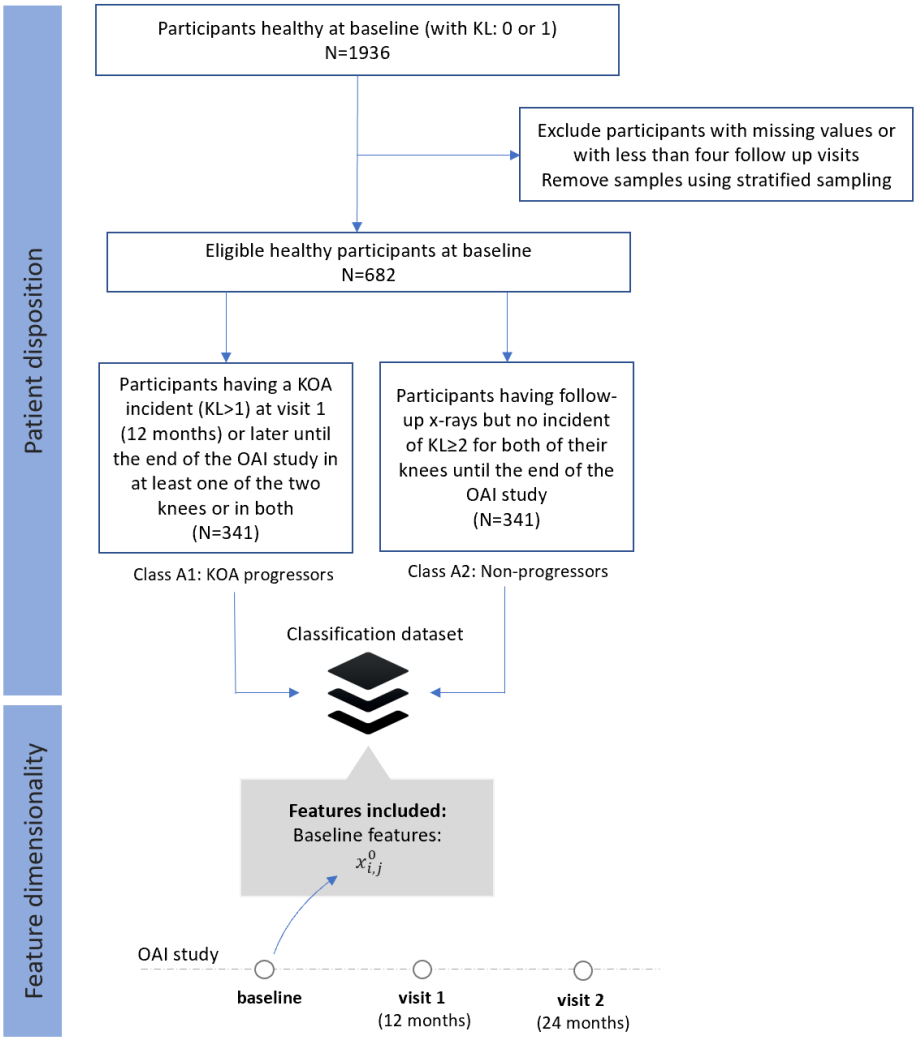


Figure 2.1. Flow chart of study design for dataset A.

- Dataset B (FS2): Progressors vs non-progressors using progression data within the first 12 months

Input: Dataset B contains data that declares the features' progression within the first 12 months. Specifically, the Equation (1) denotes the way that this progression was calculated.

$$dx^k_{i,j} = x^k_{i,j} - x^0_{i,j} \quad , \forall j \in \mathcal{F} \qquad 1. \quad (1)$$

where $x^k_{i,j}$ and $x^0_{i,j}$ are the j components (features) of sample x_i measured at the visit k and the baseline (visit 0), respectively; $dx^k_{i,j}$ is the calculated progression of $x_{i,j}$ within the time period between the k-th visit and the baseline and $\mathcal{F}$ denotes the subset of features that co-exist in both visits (233 features for dataset B). As an example, let us consider the participant x_{100} with a body mass index (*P01BMI*) of 20 at the baseline visit ($x^0_{100,49}$ = 20, where j = 49 is the index of feature *P01BMI*). Let us also assume that the participant's BMI at visit 1 has increased to 25 ($x^1_{100,49}$ = 25). Thus, the BMI

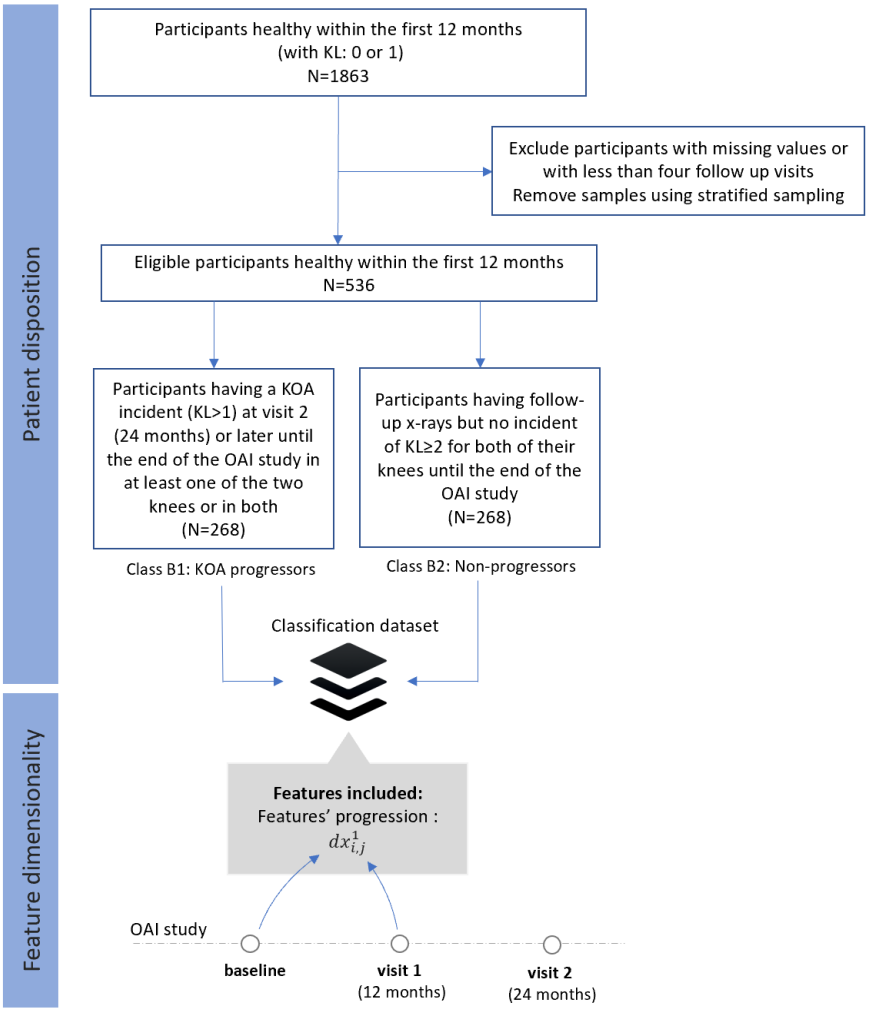
Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

progression of the specific participant is calculated as $dx^1_{100,49} = 25 - 20 = 5$. This calculation has been performed for all the 233 features of dataset B.

After data resampling, the following two classes of participants were created (Figure 2.2), as follows:

- Class B1 (KOA progressors): This class comprises progression data $dx^1_{i,j}$ of 268 participants who were healthy (KL 0 or 1) within the first 12 months (both at the baseline and the visit 1), but they had an incident of KL ≥ 2 at the second visit (24 months) or later (until the end of the OAI study).
- Class B2 (non-progressors): This class involves progression data $dx^1_{i,j}$ from 268 participants with KL 0 or 1 at the baseline, who had follow-up x-rays with no other incident of KL ≥ 2 in any of their knees until the end of the OAI study.

Output: Classification outputs 0 and 1 corresponding to assignments to classes B1 and B2, respectively.



Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Figure 2.2. Flow chart of study design for dataset B.

- Dataset C (FS3): Progressors vs non-progressors using progression data within the first 24 months

Input: Dataset C contains progression data $dx_{i,j}^2$ within the first 24 months (until visit 2). The dataset contains 275 features that co-exist in visit 2 and the baseline, whereas the same methodology was used to calculate the features as given in equation (1) using k = 2. The participants were divided into two equal categories (Figure 2.3), as follows:

- Class C1 (KOA progressors): This class comprises of 239 participants who had KL 0 or 1 during the first 24 months, whereas a KOA incident (KL ≥ 2) observed at visit 3 (36 months) or later during the OAI course in at least one of the two knees or in both.

- Class C2 (non-progressors): This class involves 239 participants with KL grade 0 or 1 at baseline, with follow-up X-rays and no further incidents (KL ≥ 2) for both of their knees.

Output: Classification outputs 0 and 1 corresponding to assignments to classes C1 and C2, respectively.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

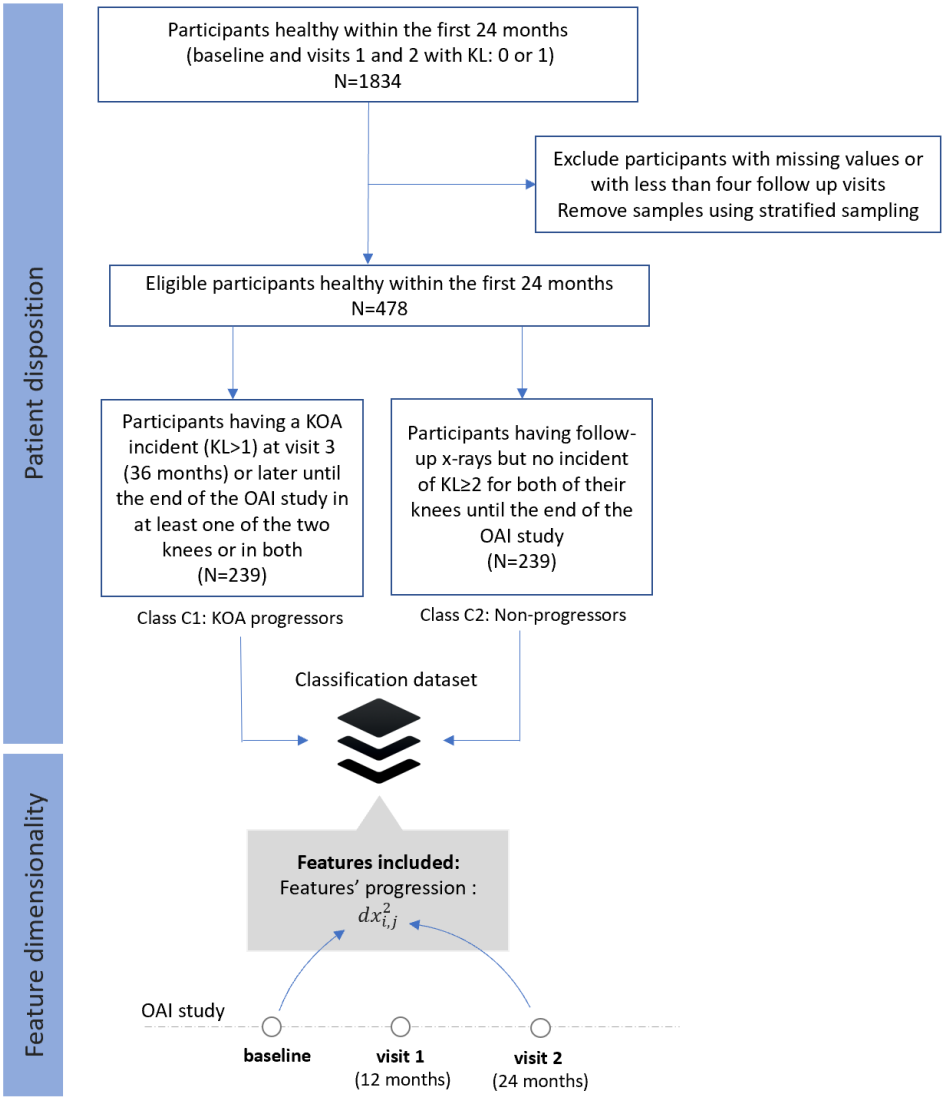


Figure 2.3. Flow chart of study design for dataset C.

- Dataset D (FS4): Progressors vs non-progressors using data from the baseline visit along with progression data within the first 12 months

Input: Dataset D contains 957 features from both datasets A and B. Specifically, it consists of 957 features from the baseline ($x^0_{i,j}$, $j = 1, \dots, 724$) along with progression data ($dx^1_{i,j}$, $j = 1, \dots, 233$) within the first 12 months. The list with the selected features from dataset D is given in the appendix A. After the application of data sampling, the participants were divided into two equal categories (Figure 2.4), as follows:

- Class D1 (KOA progression): This class comprises 270 participants (KL 0 or 1) who were heathy during the first 12 months (with no incident at the baseline and the first visit) and then they had an incident (KL ≥ 2) recorded at their second visit (24 months) or later until the end of the OAI study.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

- Class D2 (non-KOA): This class involves 270 healthy participants with KL0 or 1 at baseline with no further incidents in both of their knees until the end of the OAI data collection.

Output: Classification outputs 0 and 1 corresponding to assignments to classes D1 and D2, respectively.

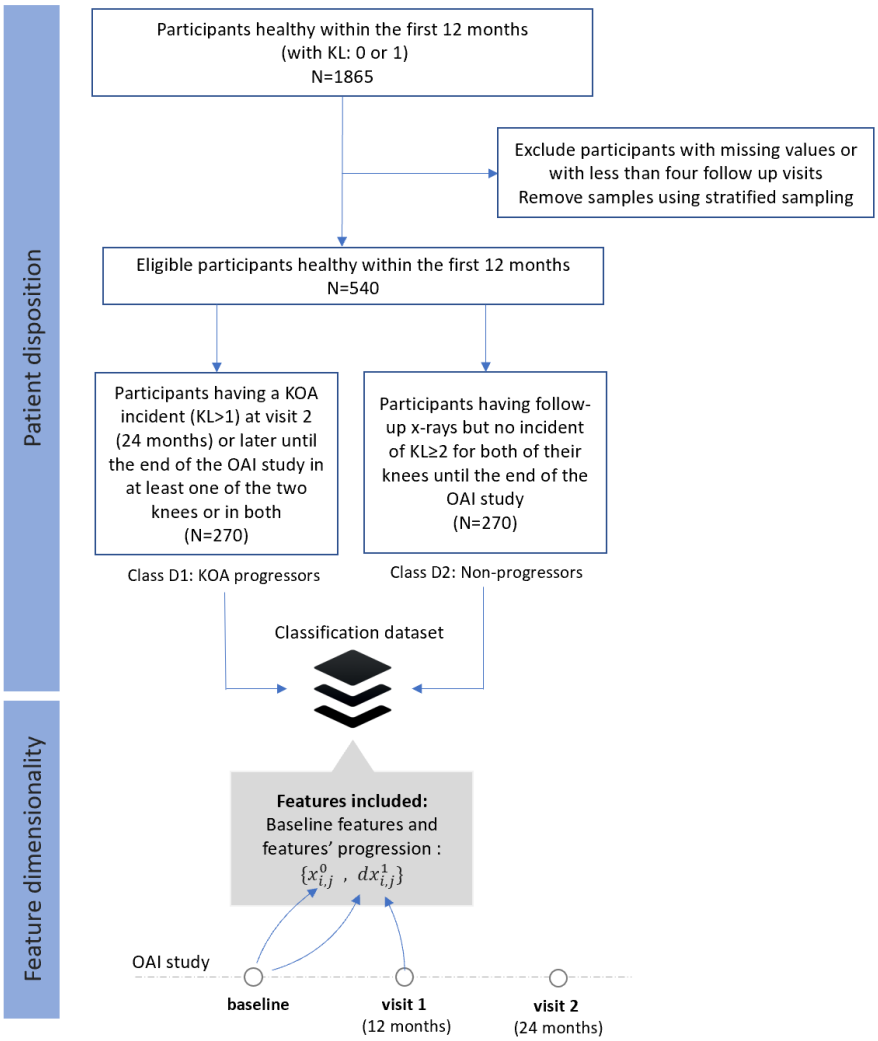


Figure 2.4. Flow chart of study design for dataset D.

- Dataset E (FS5): Progressors vs non-progressors using data from the baseline visit along with progression data within the first 24 months

Input: Dataset E contains 999 features combining datasets A and C. This set of features consists of baseline data $x^0_{i,j}$, $j = 1, \dots, 724$) as well as progression data ($dx^2_{i,j}, j = 1, \dots, 275$) within the first 24 months. Similarly, participants were divided into two equal categories (Figure 2.5), as follows:

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

- Class E1 (KOA progression): This class comprises 248 participants who were healthy (KL 0 or 1) in the first 24 months, but they had a KOA incident (KL ≥ 2) at the third visit (36 months) or later until the end of the OAI study in at least one of the two knees or in both.
- Class E2 (non-KOA): This class involves 248 healthy participants (KL0 or 1) with no further progression of KOA in both of their knees until the end of the OAI study.

Output: Classification outputs 0 and 1 corresponding to assignments to classes E1 and E2, respectively.

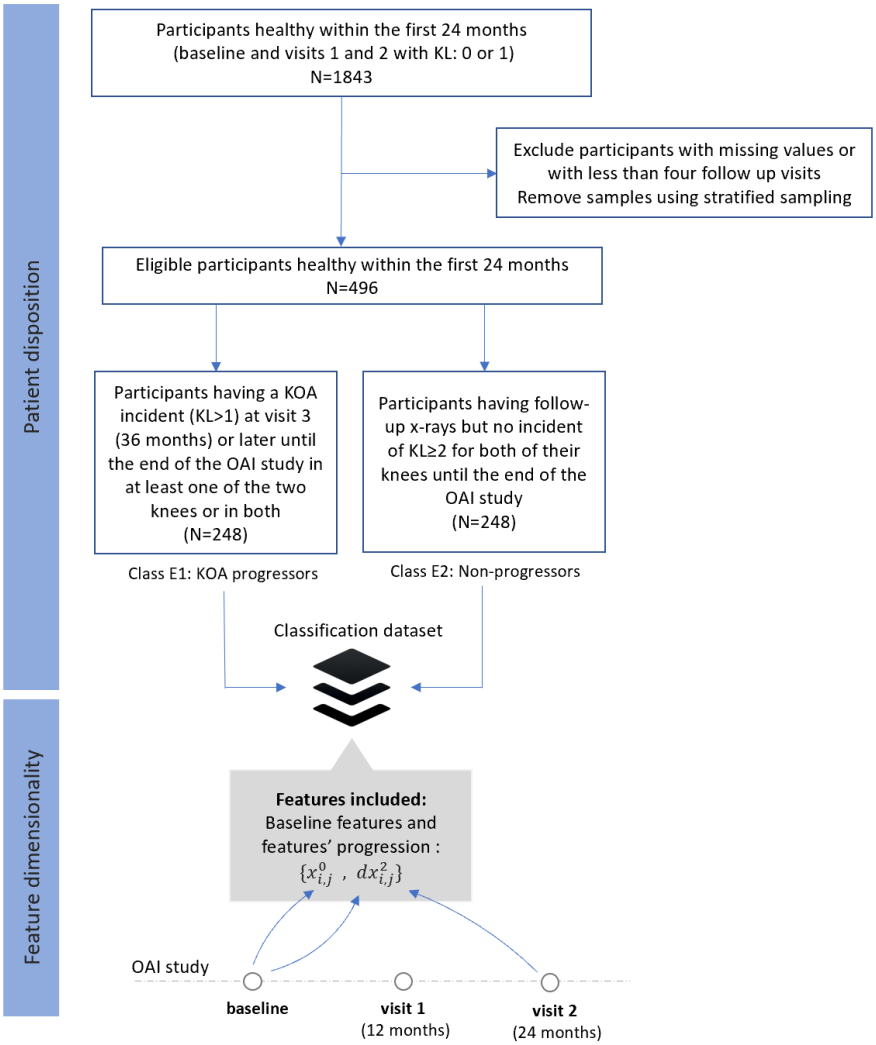


Figure 2.5. Flow chart of study design for dataset E.

Methodology

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

The proposed in this study ML methodology for KOA prediction includes four processing steps: (1) data pre-processing of the collected clinical data, (2) feature selection using the proposed approach, (3) learning process via the use of well-known ML models and (4) evaluation of the classification results. More details about the proposed methodology are presented in the following sections.

Pre-Processing

Data cleaning was initially performed by excluding the columns with more than 20% missing values compared to the total numbers of subjects. Subsequently, data imputation was performed to handle missing values. Specifically, mode imputation was implemented to replace missing values of the categorical or numerical variables by the mode (most frequent value) of the non-missing variables [146]. Standardization of a dataset is a common requirement for many ML estimators. In our work, data was normalised to [0, 1] to build a common basis for the feature selection algorithms that follow [147]. Data resampling was employed to cope with the class imbalance problem. Specifically, the majority class was reduced in order to have the same number of samples as in the minority class.

Feature Selection (FS)

A robust feature selection methodology was employed that combined the outcomes of six FS techniques: two filter algorithms (Pearson correlation [148] and Chi-2 [149]), one wrapper (with logistic regression [150]) and three embedded ones (logistic regression L2 [151], random forest [152] and LightGBM [153]). Feature ranking was decided on the basis of a majority vote scheme. Specifically, we performed all six FS techniques separately, each one resulting into a selected FS. A feature receives a vote every time it has been selected by one of the FS algorithms. We finally ranked all features with respect to the votes received.

The proposed feature selection proceeds along the following steps as shown in Figure 2.6.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Step 1: All features were normalized as described in Pre-processing Section

Step 2: We performed each one of the six FS techniques separately resulting to the creation of the following six feature subsets FS_i, i=1…,6

Step 3: Main loop

Step 3.1 For each feature j, we set V_j=0, j=1,…,M where M the total number of features

Step 3.2 Set j=1

Step 3.3 if feature j is selected in FS_i, then $V_j=V_j+1$;

Step 3.4: Repeat step 3.3 for each one of the six FS techniques for i=1,…,6

Step 3.5 Set j=j+1

Step 3.6 Terminate main loop if j>m otherwise go to step 3.3

Step 4: Rank features to descending order with respect to V_j (that is the final selection criterion)

End

Figure 2.6. Pseudocode for the implementation of the proposed feature selection (FS).

Learning Process

Various ML models were evaluated for their suitability in the task of KOA prediction. A brief description of these models is given below.

We tested logistic regression [154] which is likely the most commonly used algorithm for solving classification problems. Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1. We also evaluated decision trees (DTs) [155] which are a non-parametric supervised learning method used for classification and regression. They are simple to understand and to interpret. DTs require little data preparation and perform well even if their assumptions are somewhat violated by the true model from which the data were generated.

K-Nearest Neighbor (KNN) [156] as well as non-linear support vector machines (SVM) algorithms [116], which can deal with the overfitting problems that appear in high-dimensional spaces. In the classification setting, the KNN algorithm essentially boils

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular one is the Euclidean distance method. Furthermore, SVMs are a set of supervised learning methods used for classification, regression and outlier's detection. They are effective in high dimensional spaces and still effective in cases where the number of dimensions is greater than the number of samples.

The ensemble technique Random Forest (RF) [94] was also evaluated using DT models as weak learners. RF classifier creates a set of decision trees from randomly selected subsets of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. XGboost [157] and naive Bayes [71] algorithms were also considered. XGboost model is a sum of CART (tree) learners which try to minimize the log loss objective and the scores at leaves. These scores are actually the weights that have a meaning as a sum across all the trees of the model. Furthermore, they are always adjusted in order to minimize the loss. Moreover, naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Naive Bayes learners and classifiers can be extremely fast. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution.

Hyperparameter selection was implemented to optimize the performance of our models and to avoid overfitting and bias errors. Each model was optimized with respect to a number of preselected hyperparameters (Table 2.2). Specifically (i) 'gamma': [0,0.4,0.5,0.6], 'maximal depth': [1,2,3,4,5,6,7,8], 'minimum child and weight': [1,3,4,5,6,8] were optimized for XGboost, (ii) 'criterion': ['gini', 'entropy'], 'minimum samples leaf': [1,2,3], 'minimum samples split': [3,4,5,6,7] and 'number of estimators': [10,15,20,25,30] for random forest, (iii) 'maximal features': ['auto', 'sqrt', 'log2'], 'minimum samples leafs': [1,2,3,4,5,6,7,8,9,10,11] and 'minimum number of decision splits': [2,3,4,5,6,7,8,9,10,11,12,13,14,15] for decision trees, (iv) 'C': [0.001,0.01,0.1,1,2,3,4,5,6,7,8,9,10] and 'kernel': ['linear','sigmoid','rbf','poly'] for SVMs, (v) 'k-parameter': [5,7,9,12,14,15,16,17] for KNN and (vi) 'penalty': ['l1', 'l2'] and 'C': [100, 10, 1.0, 0.1, 0.01] for logistic regression.

Table 2.2. Hyperparameters description.

ML Models	Hyperparameters	Description
XGboost	Gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Random Forest	Maximal depth	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
	Minimum child and Weight	Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning.
	Criterion	The function to measure the quality of a split.
	Minimum samples leaf	The minimum number of samples required to be at a leaf node.
	Number of estimators	The number of trees in the forest.
	Maximal features	The number of features to consider when looking for the best split.
Decision Trees	Minimum samples split	The minimum number of samples required to split an internal node
	Minimum number of leafs	The minimum number of samples required to be at a leaf node.
SVMs	C	Regularization parameter. The strength of the regularization is inversely proportional to C.
	Kernel	Specifies the kernel type to be used in the algorithm.
KNN	k-parameter	Number of neighbors to use by default for k neighbors queries.
Logistic Regression	Penalty	Used to specify the norm used in the penalization.
	C	Inverse of regularization strength; must be a positive float.

Validation

A hold out 70–30% random data split was applied to generate the training and testing subsets, respectively. Learning of the ML was performed on the stratified version of the training sets and the final performance was estimated on the testing sets. We also evaluated the classifiers performance in terms of the confusion matrix as an additional evaluation criterion.

Confusion matrix is a way to evaluate the performance of a classifier. Specifically, a confusion matrix is a summary of prediction results on a classification problem (Table 2.3). To be created the confusion matrix, the number of correct (true) and incorrect (false) predictions are summarized with count values and broken down by each class.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Table 2.3. Confusion matrix.

		Actual Classes	
		Positive	Negative
Predicted classes	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Results

In this section, we present the most important risk factors as they have been selected by the proposed hybrid FS methodology. Moreover, the overall performance of the models is presented in relation to the number of selected features and then reference is made to the models with the highest accuracies. Results are initially given per dataset and an overall assessment is provided at the end. The efficacy of the proposed FS methodology is also compared with the performance of the six individual FS criteria.

Prediction Performance

The proposed ML methodology was applied on each of the five datasets. Specifically, the proposed FS was executed on the pre-processed versions of the datasets ranking the available features with respect to their relevance with the progression of OA. Then the proposed ML models were trained on feature subsets of increasing dimensionality (with a step of 5). These feature subsets were generated by sorting the features according to the selected ranking. This means that the proposed ML models were trained to classify KOA progressors and non-progressors based on the first (5, 10, 15, etc.) most informative features and the testing classification accuracies were finally calculated until the full feature set has been tested. The classification results on the five datasets are given below.

- **Dataset A**

Figure 2.7 depicts the testing performance (%) of the competing ML models with respect to the number of selected features for dataset A. In particular, DTs failed in this task, recording low testing performances (in the range of 42.44–65.85%). In contrast, the other models had an upward trend in the first 20–60 features, followed by a steady testing performance in most of the cases. Specifically, the logistic regression model

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

showed an upward trend with respect to selected features in the first 30–50 features, with a maximum of 71.71% at 50 features (which was the overall best performer). The inclusion of additional features led to a small reduction in the accuracies achieved.

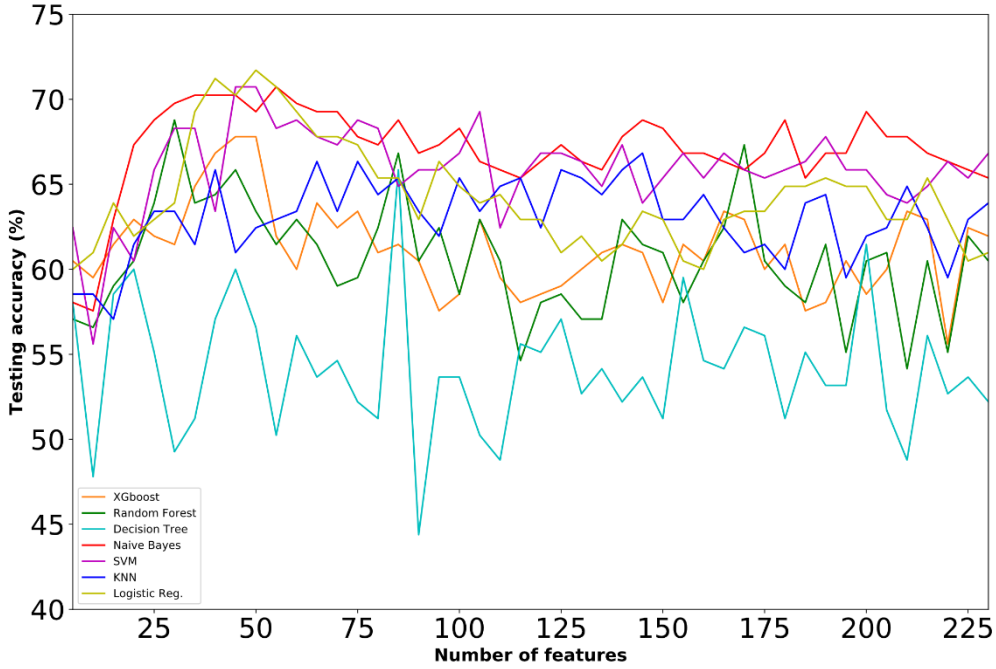


Figure 2.7. Learning curves with testing accuracy scores on dataset A for different machine learning (ML) models trained on feature subsets of increasing dimensionality.

Table 2.4 summarizes the results of logistic regression, XGboost, SVM, random forest, KNN, naive Bayes and DT on the two-class problem. A moderate number of features (in the range of 30–55) was finally selected by the majority of the ML models (in five out of the seven), whereas the overall maximum was achieved by LR on a group of fifty selected (50) risk factors. KNN and DTs selected more features (145 and 85, respectively) leading to low accuracies. The second highest accuracy was received for SVM and Naive Bayes (70.73% in both), whereas lower accuracies were obtained by NB, RF and XGboost.

Table 2.4. Best testing accuracies achieved for ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. A1 and A2 denote classes 1 and 2 of dataset A, respectively.

Models	Accuracy (%)	Confusion Matrix	Features	Parameters
	71.71	A1 A2	50	Penalty: l1, C: 1.0

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Logistic Regression		A1	73	28		
		A2	30	74		
			A1	A2		
Naive Bayes	70.73	A1	72	29	55	GaussianNB
		A2	31	73		
			A1	A2		
SVM	70.73	A1	75	26	45	C = 2, kernel = sigmoid
		A2	34	70		
			A1	A2		
KNN	66.83	A1	78	23	145	leaf_size: 1, n_neighbors: 12, weights: distance
		A2	45	59		
			A1	A2		
Decision Tree	65.85	A1	68	33	85	max_features: log2, min_samples_leaf: 4, min_samples_split: 11
		A2	37	67		
			A1	A2		
Random Forest	68.78	A1	71	30	30	criterion: gini, min_samples_leaf: 3, min_samples_split: 7, n_estimators: 15
		A2	34	70		
			A1	A2		
XGboost	67.8	A1	69	32	45	gamma: 0, max_depth: 1, min_child_weight: 4
		A2	34	70		

- **Dataset B**

Figure 2.8 demonstrates the testing performance (%) of the competing ML models with respect to the number of selected features for dataset B. The following remarks could be extracted from Figure 2.8: (i) Considerably lower accuracies were achieved by all the competing ML models compared to the ones received in dataset A; (ii) LR and NB gave the maximum testing performance of approximately 64% at 25 features (which was the overall best performer in dataset B). The addition of more features did not increase the testing performance of the model but led to a reduction in the accuracies achieved. (iii) Low testing performances were accomplished by the rest of the ML models (in the range of 42.24–62.11%). The accuracies and confusion matrixes reported

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

in Table 2.5 verify the aforementioned results. In all the competing models, the best accuracies were recorded using a relatively small number of selected risk factors (less or equal to 40).

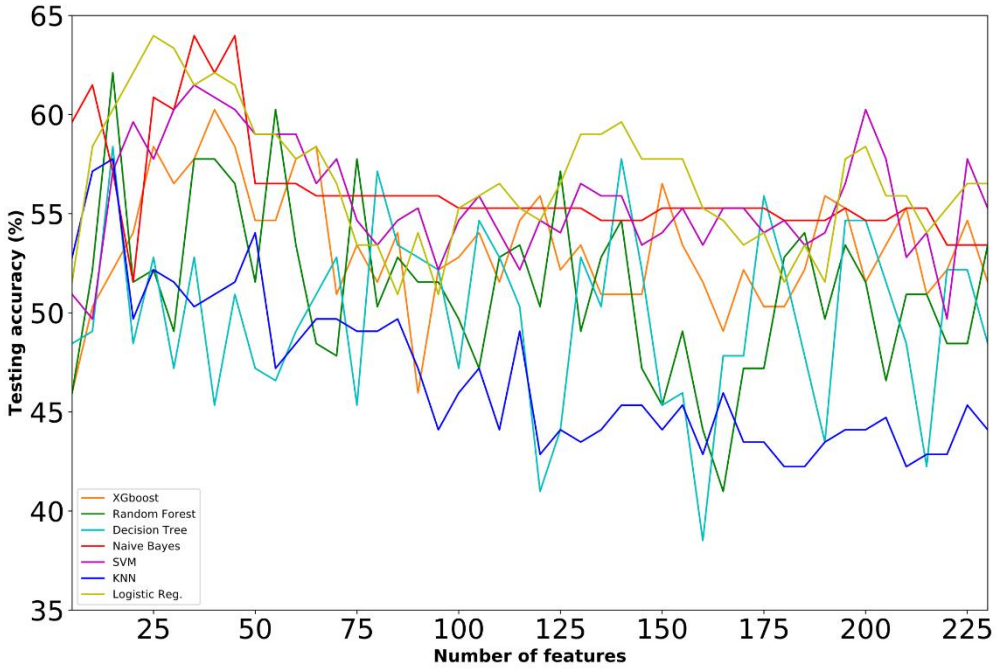


Figure 2.8. Learning curves with testing accuracy scores on dataset B for different ML models trained on feature subsets of increasing dimensionality.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Table 2.5. Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. B1 and B2 denote classes 1 and 2 of dataset B, respectively.

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
			B1	B2		
Logistic Regression	63.98	B1	48	22	25	Penalty: l1, C: 1.0
		B2	36	55		
			B1	B2		
Naive Bayes	63.98	B1	50	20	35	GaussianNB
		B2	38	53		
			B1	B2		
SVM	61.49	B1	46	24	35	C: 6, kernel: linear
		B2	38	53		
			B1	B2		
KNN	57.76	B1	63	7	15	leaf_size: 1, n_neighbors: 16, weights: uniform
		B2	61	30		
			B1	B2		
Decision Tree	58.39	B1	41	29	15	max_features: auto, min_samples_leaf: 1, min_samples_split: 6
		B2	38	53		
			B1	B2		
Random Forest	62.11	B1	48	22	15	criterion: gini, min_samples_leaf: 2, min_samples_split: 7, n_estimators: 30
		B2	39	52		
			B1	B2		
XGboost	60.25	B1	44	26	40	gamma: 0.4, max_depth: 7, min_child_weight: 5
		B2	38	53		

- **Dataset C**

Less informative features with small generalization capacity are contained in dataset C, as reported in Figure 2.9 and Table 2.6. Unlike the previous two datasets, the best

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

testing performance for dataset C was received at 225 features using DTs (66.67%). In general, unstable and low testing performances were observed for the majority of the employed ML models. The second highest accuracy was received for SVM (65.28%), whereas lower accuracies were obtained by the rest of the models. A significant number of features (more than 100) was also required in five out of the seven FS approaches highlighting the inability of dataset C features to provide useful information for the progression of KOA.

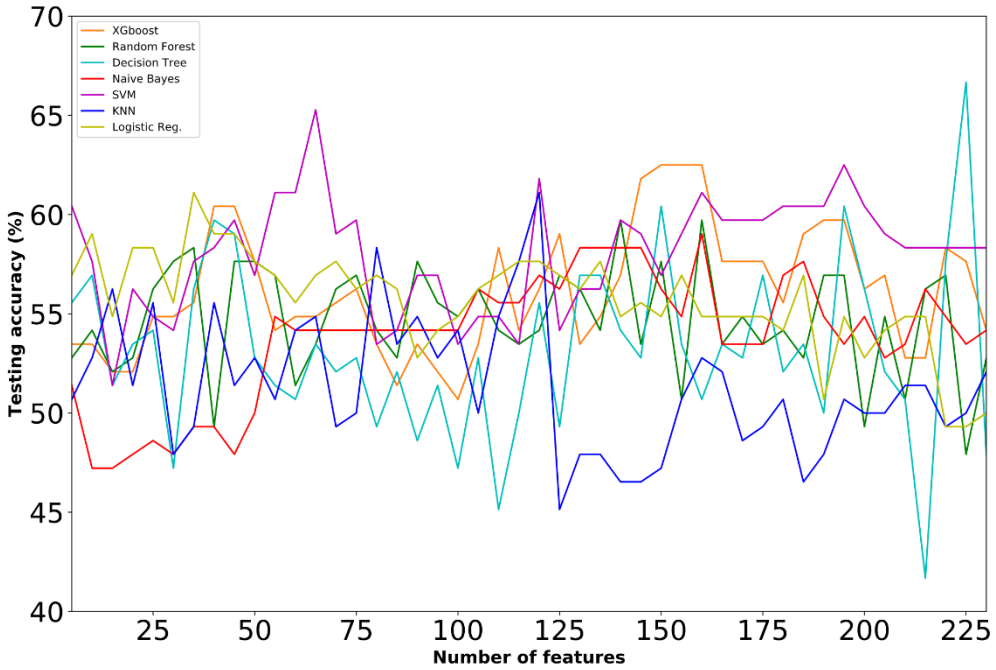


Figure 2.9. Learning curves with testing accuracy scores on dataset C for different ML models trained on feature subsets of increasing dimensionality.

Table 2.6. Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. C1 and C2 denote classes 1 and 2 of dataset C, respectively.

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	61.11		C1	C2	35	Penalty: l1, C: 1.0
		C1	49	15		
		C2	41	39		
Naive Bayes	59.03		C1	C2	160	GaussianNB
		C1	23	41		
		C2	18	62		

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

			C1	C2		
SVM	65.28	C1	48	16	65	C: 5, kernel: rbf
		C2	34	46		
			C1	C2		
KNN	61.11	C1	55	9	120	leaf_size: 1, n_neighbors: 5, weights: uniform
		C2	47	33		
			C1	C2		
Decision Tree	66.67	C1	44	20	225	max_features: auto, min_samples_leaf: 2, min_samples_split: 8
		C2	28	52		
			C1	C2		
Random Forest	59.72	C1	37	27	140	criterion: gini, min_samples_leaf': 1, min_samples_split: 5, n_estimators: 25
		C2	31	49		
			C1	C2		
XGboost	62.5	C1	44	20	150	n_estimators = 100, max_depth = 8, learning_rate = 0.1, subsample = 0.5
		C2	34	46		

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

- **Dataset D**

The combination of datasets A and B proved to be beneficial in the task of predicting KOA progression. Specifically, the following conclusions are drawn from the results reported in Figure 2.10 and Table 2.7: (i) The best performance (74.07%) was achieved by the SVM on the group of the fifty-five selected risk factors with linear kernel penalty and C = 0.1 (Dataset D). This performance was the overall best one achieved in all five datasets. (ii) The second highest accuracy was received for the logistic regression (72.84%), whereas lower accuracies were obtained by the rest of the models. (iii) SVM and LR followed a similar progression in the reported accuracies with respect to the number of selected features with an upward trend in the first 20–55 features, followed by a slight performance decrease as the number of features increases. (iv) KNN gave moderate results with a maximum testing performance of 71.6% at 75 selected features. (v) Low testing accuracies were obtained by RF, XGboost and DT in the range of 42.59–66.67%.

- **Dataset E**

In dataset E, the SVM-based approach exhibited an upward trend with respect to selected features in the first 20–70 features, with a maximum of 71.81% at 70 features (which was the best in the category). The inclusion of additional features led to a small reduction in the accuracies achieved (Figure 2.11). Similarly to SVM, LR gave the second highest accuracy (71.14%) for less features (55). XGboost also gave a comparable performance (70.47%) in a subset of 45 selected features. Lower testing accuracies were received by the rest of ML models (Table 2.8).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

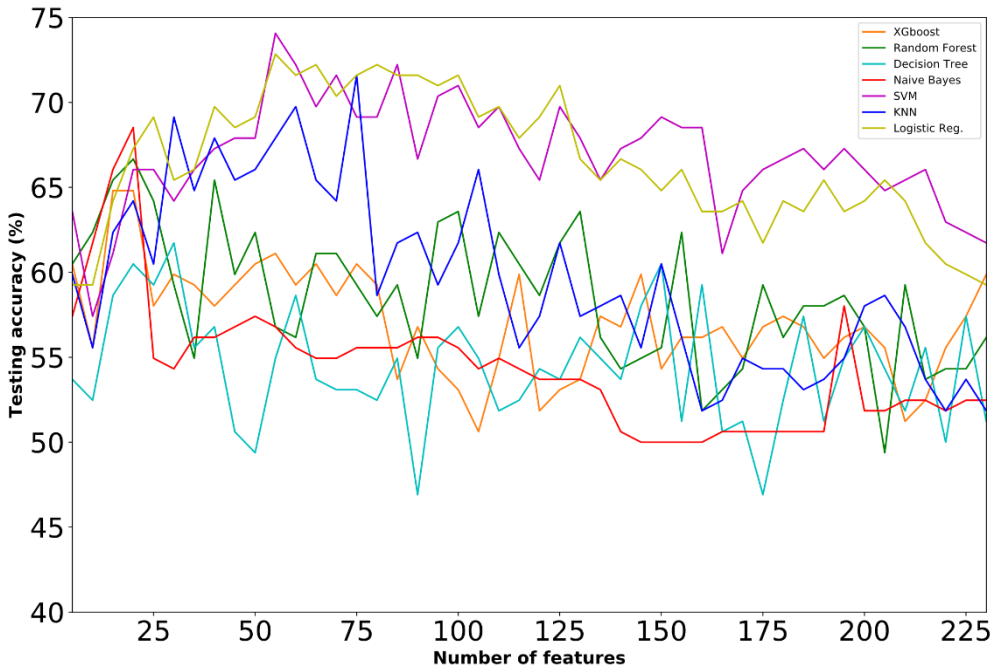


Figure 2.10. Learning curves with testing accuracy scores on dataset D for different ML models trained on feature subsets of increasing dimensionality.

Table 2.7. Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. D1 and D2 denote classes 1 and 2 of dataset D, respectively.

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	72.84		D1	D2	55	Penalty: l1, C: 1.0
		D1	54	27		
		D2	17	64		
Naive Bayes	68.52		D1	D2	20	GaussianNB
		D1	44	37		
		D2	14	67		
SVM	74.07		D1	D2	55	C: 0.1, kernel: linear
		D1	56	25		
		D2	17	64		
KNN	71.6		D1	D2	75	algorithm: auto, leaf_size: 1, n_neighbors: 17, weights: uniform
		D1	55	26		

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

		D2	20	61		
			D1	D2		
Decision Tree	61.73	D1	56	25	30	max_features: auto, min_samples_leaf: 3, min_samples_split: 10
		C2	37	44		
			D1	D2		
Random Forest	66.67	D1	47	34	20	criterion: gini, min_samples_leaf: 3, min_samples_split: 3, n_estimators: 25
		D2	20	61		
			D1	D2		
XGboost	64.81	D1	51	30	15	gamma: 0.6, max_depth: 1, min_child_weight: 8
		D2	27	54		

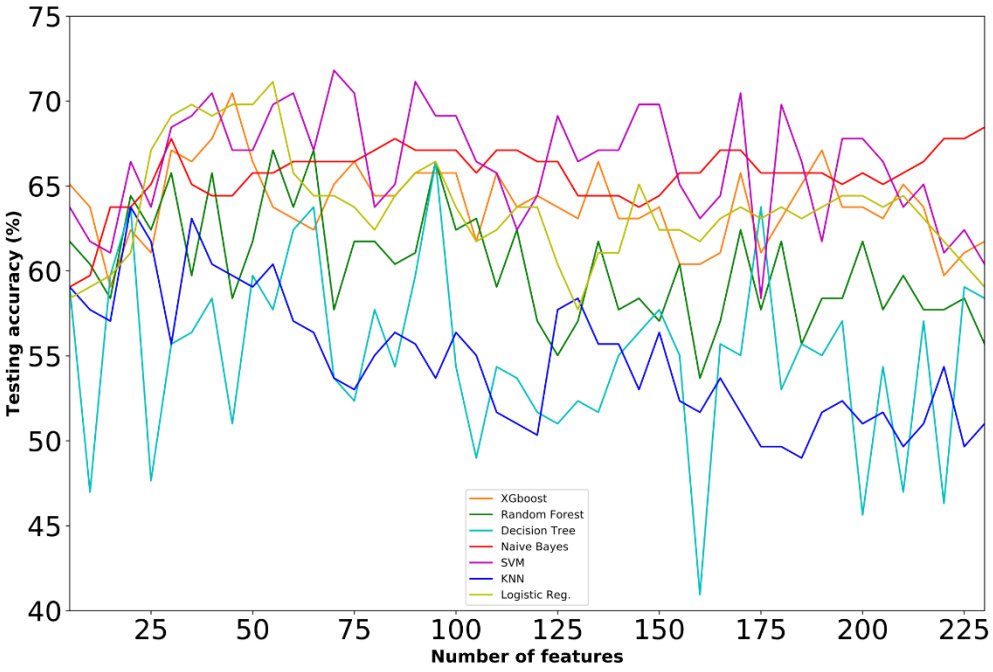


Figure 2.11. Learning curves with testing accuracy scores on dataset E for different ML models trained on feature subsets of increasing dimensionality.

Table 2.8. Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. E1 and E2 denote classes 1 and 2 of dataset E, respectively.

Models	Accuracy (%)	Confusion Matrix			Features	Parameters

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

			E1	E2		
Logistic Regression	71.14	E1	50	17	55	Penalty: l1, C: 1.0
		E2	26	56		
			E1	E2		
Naive Bayes	68.46	E1	48	19	230	GaussianNB
		E2	28	54		
			E1	E2		
SVM	71.81	E1	50	17	70	C: 1, kernel: sigmoid
		E2	25	57		
			E1	E2		
KNN	63.76	E1	48	19	20	algorithm: auto, leaf_size: 1, n_neighbors: 16, weights: uniform
		E2	35	47		
			E1	E2		
Decision Tree	66.44	E1	45	22	95	max_features: auto, min_samples_leaf: 2, min_samples_split: 12
		E2	28	54		
			E1	E2		
Random Forest	67.11	E1	42	25	55	criterion: gini, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 30
		E2	24	58		
			E1	E2		
Xgboost	70.47	E1	43	24	45	gamma: 0.6, max_depth: 2, min_child_weight: 1
		E2	20	62		

Table 2.9 cites the best accuracies achieved in each of the five datasets. The combined effect of baseline features (dataset A) and progression data $dx_{i,j}^{1}$ (dataset B) had a positive effect on the prediction capacity of the proposed methodology, as clearly shown in Table 2.7 where the testing accuracy in dataset D is increased by 2.36% compared to the result obtained in dataset A. A minor difference (0.1%) is observed on the accuracies reported for datasets A and E, demonstrating that $dx_{i,j}^{2}$ progression data have a negligible effect on the predictive capacity of the proposed methodology

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

and therefore could be omitted. The accuracies received in datasets B and C reveal that the baseline features are crucial for predicting KOA progression.

Table 2.9. Summary of all reported results.

Dataset	Data Used in the Training			Best Testing Performance (%) Achieved	Best Model	Num. of Selected Features
	Baseline	M12 Progress Wrt Beseline	M24 Progress Wrt Baseline			
A	•			71.71	Logistic Regression	50
B		•		63.98	Logistic Regression	25
C			•	66.67	Decision Tree	225
D	•	•		74.07	SVM	55
E	•		•	71.81	SVM	70

Selected Features

Figure 2.12 shows the first 70 features selected by the proposed FS approach for datasets A to E. Features are visualised with different colors and marks depending on the feature category they belong. The following conclusions could be drawn from the analysis of Figure 2.12: (i) Symptoms and medical imaging outcomes seem to be the most informative feature categories in dataset D in which the overall best performance was achieved. Specifically, eleven medical history outcomes and ten symptoms were selected in the first 55 features that gave the optimum prediction accuracy; (ii) nutrition and medical history characteristics were also proved to be contributing risk factors since approximately 20 out of the first selected 55 features were from these two feature categories (in dataset D). The full list of selected features for dataset D is provided in the appendix A; (iii) similar results with respect to the selected features were extracted from the analyses in datasets A and E (in Figure 2.12a, e) that gave comparative prediction results (close to 72%); (iv) a different order in the selected features was observed in datasets B and C (as depicted in Figure 2.12b, c). The low accuracies recorded in these datasets (less than 67%) verify that the contained in these datasets features are less informative; (v) overall, it was concluded that a combination of heterogeneous features coming from almost all feature categories is needed to

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

predict KL progression highlighting the necessity of adopting a multi-parametric approach that could handle the complexity of the available data.

Comparative Analysis

To evaluate the effectiveness of the proposed FS methodology, a comparison was performed in this section between the hybrid FS mechanism and the six well known FS techniques (the ones that are contained within the selection mechanism of the proposed methodology). The comparison was performed on dataset D that gave the overall best prediction performance. SVM was finally used to evaluate the prediction capacity of all the FS techniques considered here.

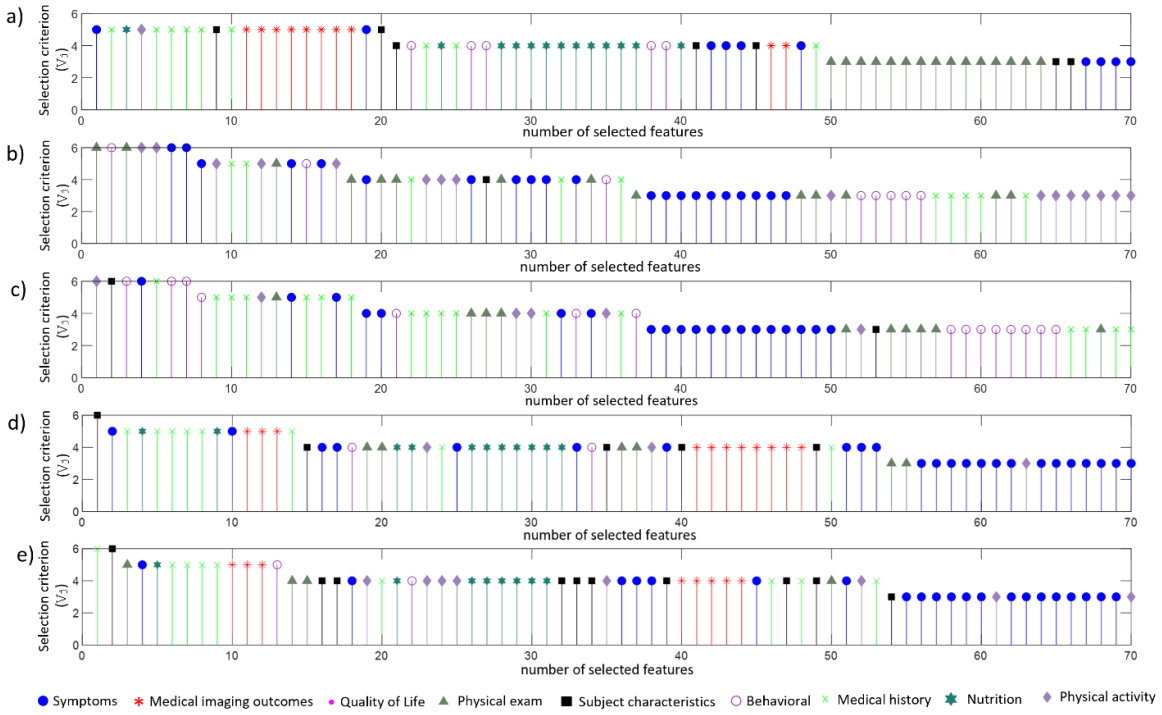


Figure 2.12. Features selected in datasets A to E in (**a–e**), respectively. Axis y (selection criterion) denotes how many times a feature has been selected (6 declares that a specific feature has been selected by all six FS techniques and so on). Features have been ranked based on the selection criterion V_j and are visualised with different colors each one representing a specific feature category.

We performed and validated all six FS techniques separately, each one resulting into a different feature subset. SVM was finally trained on the resulted feature spaces of increasing dimensionality and the optimum feature subset was identified per case. As indicated in Table 2.10, the majority of the competing FS techniques provided lower testing performances compared to the proposed FS methodology. The wrapper

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

technique based on LR was the only one that achieved an equal testing performance (%) with the proposed FS methodology. Specifically, the wrapper FS achieved its maximum accuracy at 70 features, while the proposed FS methodology achieved the same accuracy score using a smaller feature subset (55 features).

Table 2.10. Testing performance (%) of the competing FS techniques with respect to the number of selected features for dataset D.

	FS Criteria						
	Filter Algorithms		Wrapper Algorithms	Embedded			Proposed FS Criterion
Features	Chi-2	Pearson	Logistic Regression	Logistic Regression (L2)	Random Forest	LightGBM	
5	58.02	62.35	62.96	54.32	45.68	56.17	63.58
10	63.58	63.58	59.88	51.23	48.77	50.00	57.41
15	61.11	58.02	51.85	50.62	50.62	53.70	61.11
20	53.09	61.11	57.41	48.77	50.62	50.00	66.05
25	60.49	65.43	60.49	51.85	56.79	53.70	66.05
30	64.81	70.37	70.37	60.49	58.02	51.23	64.2
35	66.67	65.43	62.96	56.79	58.02	53.70	66.05
40	59.26	66.67	65.43	60.49	60.49	54.32	67.28
45	64.81	67.90	69.75	54.32	58.02	46.30	67.9
50	63.58	67.28	68.52	55.56	60.49	48.77	67.9
55	64.81	69.75	64.81	53.09	59.88	53.09	74.07
60	69.75	67.28	65.43	55.56	59.88	55.56	72.22
65	61.73	64.81	70.99	60.49	58.64	54.94	69.75
70	68.52	66.67	74.07	56.17	56.17	54.32	71.6
75	68.52	64.81	72.22	54.32	51.85	59.26	69.14
80	66.05	66.67	69.14	58.02	58.02	59.88	69.14
85	66.05	66.67	72.84	53.70	59.26	57.41	72.22
90	67.90	56.79	73.46	58.64	62.96	53.09	66.67
95	66.67	56.79	69.14	59.88	61.11	55.56	70.37
100	62.96	59.88	72.22	61.73	56.79	55.56	70.99

Discussion

This work focuses on the development of a ML-empowered methodology for KL grades prediction in healthy participants. The prediction task has been coped as a two-class classification problem where the participants of the study were divided into two

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

groups (KOA progressors and non-progressors). Various ML models were employed to perform the binary classification task (KOA progressors versus non-progressors) where accuracies up to 74.07% (Dataset D) were achieved. Within the secondary objectives of the study were to identify informative risk factors from a big pool of available features that contribute more to the classification output (KOA prediction). Moreover, we explored different options with respect to the time period within which data should be considered in order to reliably predict KOA progression.

Three different options were investigated as far as the time period within which data should be considered in order to reliably predict KOA progression. To accomplish this, we worked with 5 different datasets. We first examined whether baseline data (dataset A) could solely contribute in predicting KOA progression. Going one step further, the features 'progression within the first 12 months or 24 months was also considered as an alternative source of information (datasets B and C). The aforementioned analysis revealed that: (i) a 71.71% prediction performance can be achieved using features from the baseline, (ii) features' progression cannot solely provide reliable KOA predictions and (iii) a combination of features is required to maximize the prediction capability of the proposed methodology. Specifically, the overall best accuracy (74.07%) was obtained by combining datasets A and B that contain features from the baseline visit along with their progression over the next 12 months. Considering a longer period of time (24 months) in the calculation of features' progression resulted to lower prediction accuracies (71.81%).

The proposed FS methodology outperformed six well-known FS techniques achieving the best tradeoff between prediction accuracy and dimensionality reduction. From the pool of approximately 700 features of the OAI dataset, fifty-five were finally selected in this work to predict KOA. As far as the nature of the selected features, it was concluded that symptoms, medical imaging outcomes, nutrition and medical history are the most important risk factors contributing considerably to the KOA prediction. However, it was also extracted that a combination of heterogeneous features coming from almost all feature categories is needed to effectively predict KL progression.

Seven ML algorithms were evaluated for their suitability in implementing the prediction task. Table 2.7 with the summary of all reporting result indicates that LR and SVM were proved to be the best performing models. The good performance of SVM could be attributed to the fact that SVM models are particularly well suited for classifying small or medium-sized complex datasets (both in terms of data size and dimensionality). LR was the second-best performer providing the highest prediction accuracy in datasets A and B and the second highest in datasets D and E. The fact that a generalized linear model such as LR accomplishes high performances indicates that the power of the proposed methodology lies on the effective and robust mechanism of selecting important risk factors and not so much on the complexity of the finally

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

employed classifier. Identifying important features from the pool of heterogeneous health-related parameters (including anthropometrics, medical history, exams, medical outcomes, etc.) that are available nowadays is a key to increase our understanding of the KOA progression and therefore to provide robust prediction tools.

A few studies have recently addressed the problem of predicting KOA progression from different perspectives and employing different data sources. A weighted neighbor distance classifier was presented by Ashinsky et al. to classify isolated T2 maps for the progression to symptomatic OA with 75% accuracy [72]. Progression to clinical OA was defined by the development of symptoms as quantified by the WOMAC questionnaire 3 years after baseline evaluation. MRI images and PCA were employed by Du et al. to predict the progression of KOA using four ML techniques [73]. For KL grade prediction, the best performance was achieved by ANN with AUC = 0.761 and F-measure = 0.714. An MRI-based ML methodology has been also proposed by Marques et al. to prognose tibial cartilage loss via quantification of tibia trabecular bone where a odds ratio of 3.9 (95% confidence interval: 2.4–6.5) was achieved [70]. X-ray combined with pain scores have been utilized by Halilaj et al. to predict the progression of joint space narrowing (AUC = 0.86 using data from two visits spanning a year) and pain (AUC = 0.95 using data from a single visit) [79]. Similarly, another two studies (Tiulpin et al. [78] and Widera et al. [76]) made use of Xray images along with clinical data to predict KOA progression using either CNN or ML approaches achieving less accurate results. The current study is the only one employing exclusively clinical non-imaging data and also contributes to the identification of important risk factors from a big pool of available features. The proposed methodology achieved comparable results with studies predicting KL grades progression demonstrating its uniqueness in facilitating prognosis of KOA progression with a less complicated ML methodology (without the need of big imaging data and image-based deep learning networks).

Among the limitations of the current study is the relatively large number of features (55) that were finally selected as possible predictors of KOA. The selected features come from almost all feature categories highlighting the necessity of adopting a rigorous data collection process in order to formulate the input feature vector that is needed for the ML training. Moreover, the ML models employed are opaque (black boxes) and therefore they are insufficient to provide explanations on the decisions (inability to explain how a certain output has been drawn). To overcome the aforementioned challenges, it is important for AI developers to build transparency into their algorithms and/or enhance the explainability of existing ML or DL networks.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Conclusions

This work focuses on the development of a ML-based methodology capable of (i) predicting KOA progression (and specifically KL grades progression) and (ii) identifying important risk factors which contribute to the prediction of KOA. The proposed FS methodology combines well-known approaches including filter, wrapper and embedded techniques whereas feature ranking is decided on the basis of a majority vote scheme to avoid bias. Finally, a variety of ML models were built on the selected features to implement the KOA prediction task (treated as a two-class classification problem where a participant is classified to either the class of KOA progressors or to the non-progressors' class). Apart from the selection of important risk factors, this study also explores three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. The nature of the selected features was also discussed to increase our understanding of their effect on the KOA progression. After an extensive experimentation, a 74.07% classification accuracy was achieved by SVM on a group of fifty-five selected risk factors (in dataset D). Understanding the contribution of risk factors is a valuable tool for creating more powerful, reliable and non-invasive prognostic tools in the hands of physicians. For our future work, we are planning to also consider image-based biomarkers and areas with valuable information derived from biomechanical data that are expected to further improve the predictive capacity of the proposed methodology. ML explainability analysis will also be considered to capture the effect of the selected features on the models' outcome.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/).

Funding

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

This research was funded by the European Community's H2020 Programme, under grant agreement Nr. 777159 (OACTIVE).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapter 3

Identifying Robust Risk Factors for Knee Osteoarthritis Progression: An Evolutionary Machine Learning Approach

Published as:

Kokkotis, C, Moustakidis, S, Baltzopoulos, V, Giakas, G, & Tsaopoulos, D (2021) Identifying Robust Risk Factors for Knee Osteoarthritis Progression: An Evolutionary Machine Learning Approach. Healthcare 9(3) 260.

Abstract

Knee osteoarthritis (KOA) is a multifactorial disease which is responsible for more than 80% of the osteoarthritis disease's total burden. KOA is heterogeneous in terms of rates of progression with several different phenotypes and a large number of risk factors, which often interact with each other. A number of modifiable and non-modifiable systemic and mechanical parameters along with comorbidities as well as pain-related factors contribute to the development of KOA. Although models exist to predict the onset of the disease or discriminate between asymptotic and OA patients, there are just a few studies in the recent literature that focused on the identification of risk factors associated with KOA progression. This study contributes to the identification of risk factors for KOA progression via a robust feature selection (FS) methodology that overcomes two crucial challenges: (i) the observed high dimensionality and heterogeneity of the available data that are obtained from the Osteoarthritis Initiative (OAI) database and (ii) a severe class imbalance problem posed by the fact that the KOA progressors class is significantly smaller than the non-progressors' class. The proposed feature selection methodology relies on a combination of evolutionary algorithms and machine learning (ML) models, leading to the selection of a relatively small feature subset of 35 risk factors that generalizes well on the whole dataset (mean accuracy of 71.25%). We investigated the effectiveness of the proposed approach in a comparative analysis with well-known FS techniques with respect to metrics related to both prediction accuracy and generalization capability. The impact of the selected risk factors on the prediction output was further investigated using SHapley Additive exPlanations (SHAP). The proposed FS methodology may contribute to the development of new, efficient risk stratification strategies and identification of risk phenotypes of each KOA patient to enable appropriate interventions.

Keywords: knee osteoarthritis prediction; feature selection; genetic algorithm; machine learning; explainability

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Introduction

Knee osteoarthritis (KOA) has a higher prevalence rate compared with other types of osteoarthritis (OA). KOA is a consequence of mechanical and biological factors. Specifically, this complex interplay includes joint integrity, genetic predisposition, biochemical processes, mechanical forces and local inflammation. At the onset of this disease, the main consequences are low quality of life due to pain, social isolation and poor psychological state. According to the literature, age, obesity and previous injuries due to sports or occupational/daily activities show a high correlation with KOA [3, 144, 158, 159]. Particular reference should be made to the specificity of this disease. Specifically, the knee osteoarthritic process is gradual, with a variation in symptom frequency, patterns and intensity [2, 160]. Despite the constant effort of the scientific community, research on KOA prediction is still necessary to investigate and explore the multifactorial nature of the disease.

One of the main challenges is the development and refinement of prognostic KOA models that will be applicable to the entire population. In this effort, an increase has been observed in the number of studies using artificial intelligence techniques due to the existence of big data [92, 145, 161-164]. As a result of this, several techniques have been reported in the literature in which feature selection (FS) techniques and machine learning (ML) models were used to predict KOA [5, 6]. There are several studies where heterogenous datasets were considered including symptoms and nutrition questionnaires, medical imaging outcomes, subject characteristics and behavioral and physical exams. Lazzarini et al. used a guided iterative feature-elimination algorithm and principal component analysis (PCA) and they demonstrated that it is possible to accurately predict the incidence of KOA in overweight and obese women using a small subset of the available information [81]. Specifically, they achieved their aim by using only five variables and Random Forest (RF) with an area under the curve (AUC) of 0.823. In another study, Du et al. used PCA and four well-known ML models to predict the change of Kellgren and Lawrence, joint space narrowing on the medial compartment and joint space narrowing on the lateral compartment grades by using magnetic resonance imaging (MRI) [73]. They demonstrated that there are more informative locations on the medial compartment than on the lateral compartment. They achieved an AUC of 0.695–0.785. Furthermore, Halilaj et al. built a model to predict long-term KOA progression taking into account self-reported knee pain, radiographic assessments of joint space narrowing from the Osteoarthritis Initiative (OAI) database and least absolute shrinkage and selection operator (LASSO) regression models [79]. In this task, an AUC of 0.86 for radiographic progression was achieved on a 10-fold cross-validation scheme. In another study, Pedoia et al. used topological data analysis as a feature engineering technique in combination with MRI

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

and biomechanics multidimensional data [75]. In the attempt to meet the existing gap in multidimensional data analysis for early prediction of cartilage lesion progression in KOA, they used logistic regression as the ML model, achieving an AUC of 0.838. Moreover, in the task of predicting KOA severity, Abedin et al. made use of elastic net regression and were able to (i) identify the variables that have high predictive power and (ii) quantify the contribution of each variable with an overall root mean square error (RMSE) of 0.97 [74].

In 2019, Nelson et al. [77] applied an innovative ML approach in order to identify key variables associated with a progression phenotype of KOA. Specifically, they combined distance-weighted discrimination algorithm, direction-projection-permutation testing and clustering methods to identify phenotypes that are potentially more responsive to interventions. Another study by Widera et al. was based on recursive feature elimination that selects the best risk factors for the prediction of KOA progression from incomplete imbalanced longitudinal data [76]. They used five ML models achieving F1 scores from 0.573 up to 0.689. Furthermore, Tiulpin et al. applied a multi-modal ML-based KOA progression prediction model which utilizes baseline characteristics, clinical data, radiographic assessments and the probabilities of KOA progression that are calculated from a deep convolutional neural network [78]. To handle the heterogeneity of the available data, they applied a gradient boosting machine classifier with an AUC of 0.79–0.82. Moreover, Kokkotis et al. presented a robust FS approach that could identify important risk factors in a KOA prediction task [163]. The novelty of this approach lies in the combination of well-known filter, wrapper and embedded techniques, whereas feature ranking is decided on the basis of a majority vote scheme to avoid bias. A 74.07% classification accuracy was achieved by support vector machines. In addition, Jamshidi et al. worked on the identification of important structural KOA progressors [165]. They used six FS models and the best classification accuracy was achieved by multi-layer perceptron (MLP, AUC = 0.88 and 0.95 for medial joint space narrowing at 48 months and Kellgren–Lawrence (KL) grade at 48 months, respectively). In another study, Wang et al. employed a long short-term memory model to predict KOA progression [166]. They used observed time series (5-year clinical data from OAI) and they predicted the KL grade with 90% accuracy. Despite all the aforementioned valuable contributions, few of the above studies have attempted to apply robust FS methodologies for the development of ML models for the prediction of KOA progression [6].

Therefore, there is still a significant knowledge gap on the contribution of clinical data on KOA progression prediction and their impact on the training of the associated ML predictive models.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Due to the multidimensional and imbalanced nature of the datasets that are publicly available for KOA, robust identification of the best features for the prediction of KOA is a challenging task. According to our knowledge, only few studies [167-169] have attempted to address the complicated interaction of the aforementioned challenges (high dimensionality and data imbalance) in biomedical datasets (but none in the area of KOA). Main examples of FS methods that were applied in various fields to overcome the imbalance problem are (a) resampling techniques [170-173], (b) ensemble learning techniques [167, 174, 175], (c) cost-sensitive learning [176, 177], (d) one-class learning [178, 179] and (e) active learning [180, 181]. Hence, to cope with the aforementioned FS challenges (high dimensionality and data imbalance), we propose an FS technique that incorporates a number of characteristics towards the identification of robust risk factors that generalize well over the whole dataset. The proposed FS methodology, termed GenWrapper in this work, is an evolutionary genetic algorithm (GA)-based wrapper technique that differentiates from the classical GA-based FS techniques in terms of the following: (i) GenWrapper applies random under-sampling at each individual solution, forcing the GA to converge to solutions (feature subsets) that generalize well regardless of the applied data sampling; (ii) It ranks features with respect to the number of times that they have been selected in all the individual solutions for the final population. The combined effect of the aforementioned GenWrapper characteristics leads to selected features that consistently work well at any possible data sample and, thus, have increased generalization capacity with respect to KOA progression. An extensive comparative analysis has been performed to prove the superiority of GenWrapper over well-known FS algorithms with respect to both prediction accuracy and generalization.

Methods

Dataset Description

Data were obtained from the Osteoarthritis Initiative (OAI) database (available upon request at https://nda.nih.gov/oai/, accessed on 18 June 2020), which include clinical evaluation data, a biospecimen repository and radiological (magnetic resonance and X-ray) images from 4796 women and men aged 45–79 years. The features considered in this work for the prediction of KL are shown in Table 3.1. The current study included clinical data from the baseline and the first follow-up visit at month 12 from all individuals being at high risk to develop KOA or without KOA. Specifically, the dataset contains 957 features from eight different feature categories, as shown in Table 3.1. In addition, our study was based on the Kellgren and Lawrence (KL) grade as the main indicator for assessing the OA clinical status of the participants. Specifically, the

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

variables "V99ERXIOA" and "V99ELXIOA" were used to assign participants into subgroups (classes) of participants whose KOA status progressed or not.

Table 3.1. Main categories of the feature subsets considered in this study. A brief description is given along with the number of features considered per category and for each of the two visits.

Category	Description	Number of Features from Baseline	Number of Features from Visit 1
Subject characteristics	Includes anthropometric parameters (Body mass index (BMI), height, etc.)	36	9
Symptoms	Questionnaire data regarding arthritis symptoms and general arthritis or health-related function and disability	120	80
Behavioral	Includes variables of participants' quality level of daily routine and social behavior	61	43
Medical history	Questionnaire results regarding a participant's arthritis-related and general health histories and medications	123	51 (only medications)
Medical imaging outcome	Medical imaging outcomes (e.g., joint space narrowing and osteophytes)	21	-
Nutrition	Block Food Frequency questionnaire	224	-
Physical activity	Questionnaire data regarding leisure activities, etc.	24	24
Physical exam	Participants' measurements, including knee and hand exams, walking tests and other performance measures	115	26
Number of features (subtotal):		724	233
Total number of features:		957	

Problem Definition

In this study, we consider KL grade prediction as a two-class classification problem. Specifically, the participants of the study were divided into two groups: (a) Non-progressors —healthy participants with KL0 or 1 at baseline with no further incidents in both of their knees until the end of the OAI data collection; (b) KOA progressors — participants who were healthy during the first 12 months (with no incident at baseline and the first visit) and then they had an incident (KL ≥ 2) recorded at their second visit (24 months) or later, until the end of the OAI study (Figure 3.1).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

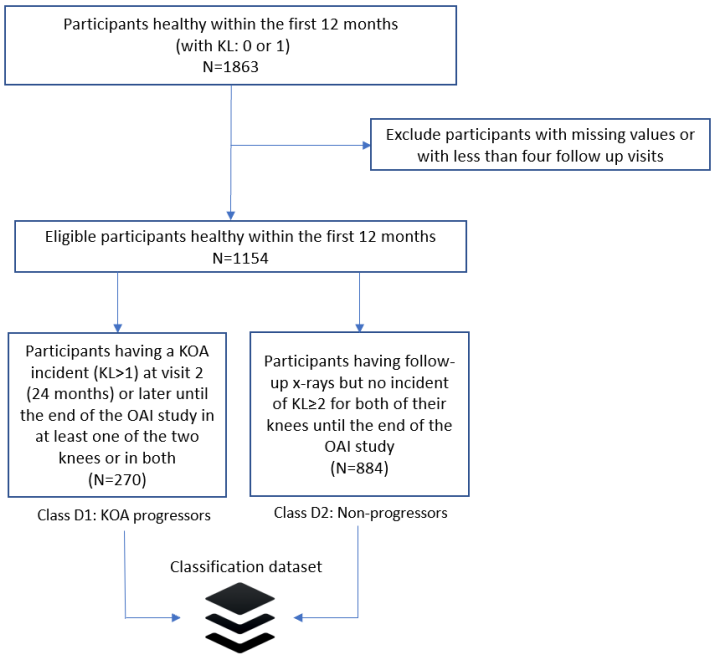


Figure 3.1. Stratification of the patients in our study and formulation of the training dataset. Inclusion/exclusion criteria are presented along with the definition of the two data classes (knee osteoarthritis (KOA) progressors and non-progressors).

Data Pre-Processing

Initially, data cleaning was performed by excluding the columns with more than 20% missing values compared to the total number of subjects. Afterwards, data imputation was performed to handle missing values. As an imputation strategy, mode imputation was implemented to replace missing values of the numerical or categorical variables by the most frequent value of the non-missing variables [147]. Standardization of a dataset is a common requirement for many ML estimators [182]. In our study, data were normalized by removing the mean and scaling to unit variance to build a common basis for the machine learning algorithms that followed. After application of the exclusion criteria, classes 1 (KOA progressors) and 2 (non-progressors) comprised 270 and 884 samples, respectively.

Feature Selection

Class imbalance is among the major challenges encountered in health-related predictive models, skewing the performance of ML algorithms and biasing predictions in favor of the majority class. To alleviate this problem, a novel evolutionary feature

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

selection is proposed in this work that overcomes the class imbalance problem and increases the generalization capacity of the finally employed ML algorithm.

The proposed FS is a genetic algorithm-based approach inspired by the procedures of natural evolution (Figure 3.2). It operates on a population of individuals (solutions), and at each generation, a new population is created by selecting individuals according to their level of fitness in the problem domain (KOA progression in our case). The individuals are then recombined using operators borrowed from natural genetics (selection, reproduction and mutation). This iterative process leads to the evolution of populations of individuals that are better suited to the problem domain. Here, each individual in the population represents an ML model trained on a specific feature subset to discriminate the aforementioned classes (KOA progressors versus non-progressors). Genes are binary values and represent the inclusion or not of particular features in the model. The number of genes is the total number of input variables in the dataset. Concatenating all genes, a so-called individual or chromosome is formulated that represents a possible solution (feature subset) in our FS problem.

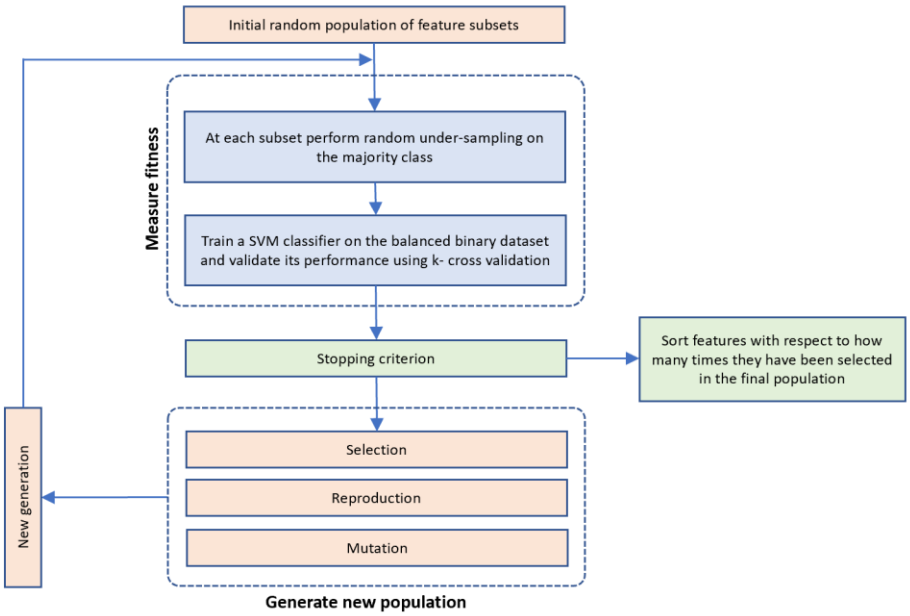


Figure 3.2. The proposed GenWrapper feature selection (FS) methodology that includes all the involved processing steps: (i) generation of the initial population; (ii) fitness measurement approach; (iii) stopping criterion; (iv) evolution mechanisms and (v) final feature ranking after the termination of the genetic algorithm (GA).

The Optimization Toolbox of MATLAB 2020b was used for the implementation of GenWrapper. The proposed FS algorithm proceeds along the following steps:

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

- Step1. Initialization

  A group of k chromosomes are randomly generated, forming the initial population of individuals.

- Step2. Fitness assignment

  A fitness value is assigned to each chromosome in the population. Specifically, the process of measuring fitness in GenWrapper can be summarized as follows. The following 3-step process (Figure 3.3) is repeated for each of the chromosomes of the population:

  Step 2.1. From the training dataset, we keep only the features that have a value of 1 in the current chromosome. This creates a truncated training set.

  Step 2.2. Random undersampling on the majority class is performed on the truncated training set. This action leads to a balanced variant of the truncated training set.

  Step 2.3. A classifier is trained on the newly produced balanced dataset. Linear support vector machines (SVMs) have been chosen as the main classification criterion due to their generalization capability.

  Step 2.4. A k-fold cross-validation scheme is employed to validate the classifier performance that is finally assigned as a fitness value to the specific individual.

- Step3. Termination condition

  The algorithm stops if the average relative change in the best fitness function value over K generations is less than or equal to a pre-determined threshold.

- Step4. Generation of a new population

  In case the termination criterion is not satisfied, a new population of individuals is generated by applying the following three GA operators:

  Selection operator: The best individuals are selected according to their fitness value.

  Crossover operator: This operator recombines the selected individuals to generate a new population.

  Mutation operator: Mutated versions of the new individuals are created by randomly changing genes in the chromosomes (e.g., by flipping a 0 to 1 and vice versa).

- Step 6. The algorithm returns to step 2.

- Step 7: Final feature ranking determination

  Upon termination of the GA algorithm, the features are ranked with respect to the number of times that they have been selected in all the individuals (chromosomes) of the final population.

  Step 7.1. A feature gets a vote when it has a value of 1 in a chromosome of the final generation.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Step 7.2. Step 7.1 is repeated for all the chromosomes of the final generation and the features' votes are summed up.

Step 7.3. Features are ranked in descending order with respect to the total number of votes received.

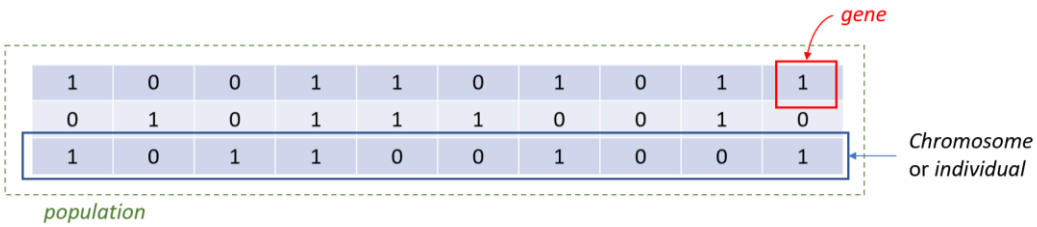


Figure 3.3. Definition of genes, chromosomes and population.

GenWrapper evaluates the fitness of each chromosome (feature subset) by firstly applying random undersampling at the associated dataset (in step 2.2) and then by training an SVM classifier on it (Figure 3.4). The k-fold cross-validation (CV) performance of the SVM is considered as the fitness of the specific individual. The best individuals (feature subsets that maximize the fitness value) are then selected and combined to generate the new population. This procedure forces the GA to converge to solutions (feature subsets) that generalize well regardless of the specific sampling that has been applied. If a specific resampling process had been applied universally on the dataset before the application of the GA-based FS, then this would lead to overfitting, since the GA algorithm would try to select the best features that fit to the specific data sample. The proposed technique integrates a random sampling mechanism when evaluating each individual, leading to features that generalize well on the whole population. Moreover, the choice of k-fold cross-validation as a validation scheme guarantees that the selected features have high predictive capacity over the whole dataset considered. Another characteristic of the proposed evolutionary FS is the way that features are selected/ranked in the final population. Instead of selecting features from the best individual in the final population, the proposed selection criterion relies on the general performance of features over the whole final population. The best solution (the one with the highest fitness value in the final population) corresponds only to the maximum possible accuracy that can be achieved by a selected feature subset on a specific subset of the whole sample. However, this does not necessarily mean that the best solution generalizes well in the whole sample. Therefore, to achieve the best possible generalization, the proposed FS ranks features with respect to the number of times that they have been selected in all the individuals of the final population. The parameters of the proposed GA-based FS have been properly selected and are cited in Table 3.2 below.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

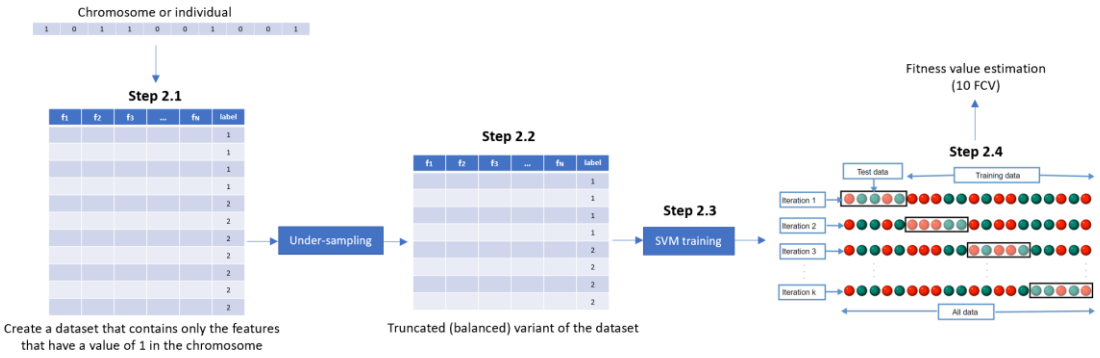


Figure 3.4. Proposed mechanism for estimating the fitness of each chromosome within a generation.

Table 3.2. Hyperparameters of the optimized GenWrapper algorithm. A brief description of each hyperparameter is provided along with the finally selected value.

Parameter	Description	Selected Value
Population size	Number of individual solutions in the population	50
Number of generations	Maximum number of generations before the algorithm halts	100
Mutation rate	Probability rate of being mutated	0.1
Crossover Fraction	The fraction of the population at the next generation, not including elite children, that the crossover function creates.	0.8
Elite Count	Positive integer specifying how many individuals in the current generation are guaranteed to survive into the next generation	5
StallGenLimit	The algorithm stops if the weighted average change in the fitness function	50
Tolerance	value over StallGenLimit generations is less than Function tolerance	1×10^{-3}1e-03

Learning

Given that the main objective of study is the identification of robust risk factors, two well-known linear ML models (linear regression (LR) and linear SVM) were utilized to evaluate the predictive capability of the selected features. The reason for employing linear models is because (i) they are computationally efficient, so they can be executed multiple times within a repetitive process such as the GA-based algorithm that is proposed in this work, and (ii) they generalize well and, therefore, can be used to assess the generalization performance of the selected features. A brief description of these models is given below.

LR is the most commonly used algorithm for solving classification problems [154]. It is an extension of the linear regression model for classification problems and it models

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

the probabilities for classification problems with two possible outcomes. SVMs are supervised learning models for classification, regression and outlier detection but are more commonly used in classification problems [116]. SVMs are effective in high-dimensional spaces and are still effective in cases where the number of dimensions is greater than the number of samples.

Validation

To evaluate the predictive capacity of the selected feature subset, a repeated cross validation process was adopted using the aforementioned classifiers. Specifically, the validation approach proceeds with the following steps

- Step 1. Random undersampling is applied on the majority class, and the retained samples along with those from the minority class form a balanced binary dataset.
- Step 2. A classifier is built on the balanced binary dataset and its accuracy is calculated using 10-fold cross-validation (10FCV).
- Step 3. Steps 1 and 2 are repeated 10 times, each one using a different randomly generated balanced dataset.
- Step 4. The final performance is calculated by averaging the obtained 10FCV classification accuracies. The resulting final performance will be referred to here as mean 10FCV.

By adopting this repeated validation approach, we guarantee that the selected features are not only suitable for a specific data sample but that they generalize well over the whole dataset. The calculated mean 10FCV performance aggregates the accuracies from 100 training runs (10 repetitions of 10FCV) on different randomly created data samples, forming a reliable measure for estimating the predictive capacity of the selected features.

Explainability

To further assess the impact of the selected features on the classification outcome, SHapley Additive exPlanations (SHAP) were considered. SHAP is a game theoretic approach that explains the output of any machine learning model and achieves the connection of the optimal credit allocation with local explanations using the classic Shapley values from game theory that come with desirable properties [183]. In this study, Kernel SHAP is used, which is a specially weighted local linear regression to

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

estimate SHAP values for any model (e.g., SVM in a two-class classification problem). The optimization of loss function L in Kernel SHAP is described below (in Equation (1)), where g is the explanation linear model that is trained on training data Z, $f(:)$ is the original prediction function to be explained and z' is a vector of 1s and 0s called coalition. Here, 1s indicate the presence of the corresponding feature, while 0 indicates its absence. $h_x(z')$ maps a feature coalition to a feature set on which the model can be evaluated, whereas $\pi_x(z')$ is the SHAP kernel.

$$(f, g, \pi_x) = \sum_{z' \in Z} \left[f\left(h_x(z')\right) - g(z') \right]^2 \pi_x(z') \quad (1)$$

Results

In this section, we demonstrate the efficiency of the proposed feature selection algorithm in comparison with other well-known FS techniques. The most significant risk factors, as selected by the proposed FS methodology, are also presented, whereas their impact on the classification result is discussed employing SHAP.

Selection Criterion

Figure 3.5 shows the evolution of the proposed fitness value with respect to the number of generations. As it was discussed, the mean fitness value is calculated by averaging the fitness values of all the 50 individual solutions in each generation. Each individual fitness value represents the performance of the employed ML model (SVM in our case) on a new, randomly generated balanced dataset (after downsampling the majority class) using k-fold cross-validation. Thus, the mean fitness value aggregates the performance of 50 employed ML models that were trained on slightly different versions of the initially available dataset. As it is observed in Figure 3.5, the mean fitness value decreases with the number of generations, meaning that the FS converges to a pool of selected feature subsets that have increased classification capacity, regardless of any specific data sampling.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

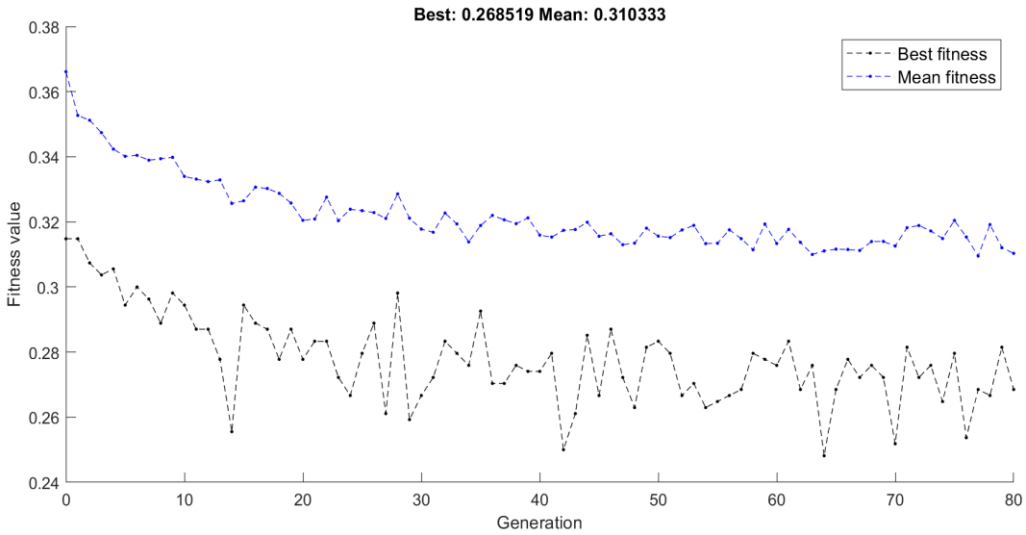


Figure 3.5. Fitness with respect to number of generations for GenWrapper. The black and blue dashed lines show the best and the mean fitness achieved at each generation, respectively.

The dashed black line in Figure 3.5 represents the minimum fitness values received at each generation of the algorithm. However, as it was noted that the best fitness value (0.26818 in our case) corresponds to a selected feature subset that has been decided based on its performance on a part of the available sample. The proposed scheme, instead of selecting the "best" feature subset of the final generation, proceeds by ranking the available features with respect to the times they have been selected in the 50 different individual solutions of the final generation. Figure 3.6 illustrates an example of such a ranking where seven features have been selected in all 50 individual solutions, another nine have been selected in 49 individual solutions and so on. The highly ranked features are the ones that are consistently selected by all individual solutions that are generated on different data samples.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

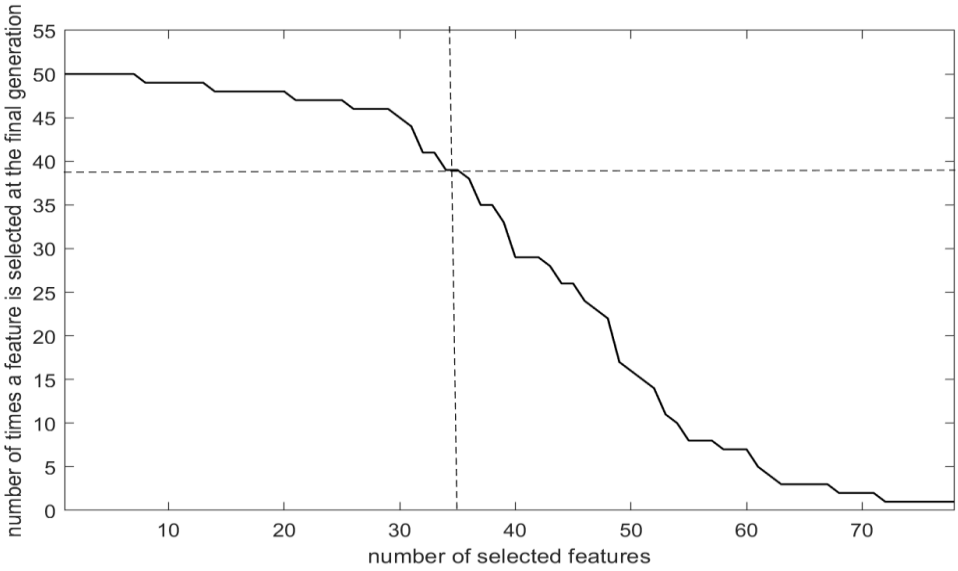


Figure 3.6. Feature ranking produced by the proposed FS (the dashed line indicates the number of features that were finally selected).

To prove the superiority of the proposed feature selection criterion over the "best" individual solution, we performed the following experimentation. Two competing feature subsets were initially extracted: (a) the proposed one that has been selected after selecting the top 35 highly ranked features and (b) the feature subset extracted from the "best" individual solution of the final GA generation (comprising 42 features). The generalization capacity of both features subsets was assessed by employing the repetitive validation approach proposed in this work and the results are shown in Table 3.3. The proposed feature ranking led to higher accuracy (in terms of mean performance, minimum and maximum accuracies), employing less features (35) compared to the ones selected in the "best" individual solution (42).

Table 3.3. Comparative analysis with respect to the final selection of features: proposed feature ranking versus the feature subset of the best individual solution in the final generation.

FS Criterion	**10FCV Accuracy Performed 10 Times**				
	Average	**Min**	**Max**	**Std**	**No. of Features**
Feature subset extracted from the "best" individual solution of the final generation	70.10%	67.59%	72.04%	1.13%	42
Proposed feature ranking	71.25%	69.22%	73.33%	1.57%	35

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Features Selected

Table 3.4 cites the 35 features selected by the chosen GenWrapper FS approach. A short description of the features and the categories in which they belong are presented. Seven out of the 35 selected risk factors come from the symptom's category, representing parameters related to pain, swelling, stiffness and knee difficulty, demonstrating the relevance of symptoms in the occurrence and progression of KOA. Moreover, eight features represent diet and nutrition-related parameters that also constitute an important risk factor category. Nine of the features are related to physical activity or exams, whereas another five behavioral risk factors were selected as relevant to KOA progression. Medical history or status estimated through subjective (three self-reported risk factors) or more objective metrics (medical imaging outcomes such as the existence of osteophytes) were also selected by the proposed FS approach. Finally, two parameters describing subject characteristics were among the selected risk factors (specifically the patient's body mass index (BMI) and height).

Table 3.4. Characteristics of the 35 most informative risk factors as selected by the proposed GenWrapper.

Selected Features	Feature Category	Description
P01BMI, P01HEIGHT	Subject characteristics	Anthropometric parameters including height and BMI
KSXRKN1, V00WOMSTFR, KPLKN1, V00WPLKN2, DIRKN16, V00KOOSYML, V00INCOME	Symptoms	Symptoms related to pain, swelling, stiffness and knee difficulty
V00EDCV, V00KQOL4, V00KQOL2, V00CESD9, CEMPLOY	Behavioral	Participants' quality level of daily routine and social behavior and social status
V00RXCHOND, V00RA, V00CHNFQCV	Medical history	Questionnaire data regarding a participant's general health histories and medications
P01SVLKOST	Medical imaging outcome	Medical imaging outcomes (e.g., osteophytes)
V00SUPCA, V00FFQ59, V00FFQSZ13, V00FFQ33, V00SUPB2, V00FFQ12, V00SUPFOL, V00FFQ19	Nutrition	Block Food Frequency questionnaire for daily average, how much each time or for past 12 months

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

PASE2, PASE6, V00PA130CV	Physical activity	Questionnaire results regarding activities during typical week or past 7 days
RKALNMT, V00lfmaxf, V00rfTHPL, V00lfTHPL, STEPST1, V00rkdefcv	Physical exam	Physical measurements of participants, including tests and other performance measures

Comparative Analysis

The performance of the proposed FS methodology was compared with eight well-known FS techniques in the recent literature. The selected techniques along with their main characteristics are briefly presented below.

A classical wrapper FS was employed in which the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It follows a time-consuming search approach by evaluating all the possible combinations of features against the evaluation criterion. The evaluation criterion is simply a performance measure which depends on the type of problem. Infinite latent feature selection (ILFS) is a probabilistic latent feature selection approach that performs the ranking step by considering all the possible subsets of features, bypassing the combinatorial problem [184]. Unsupervised graph-based filter (Inf-FS) is another FS algorithm proposed, again, by Roffo et al. (2015) [185]. In Inf-FS, each feature is a node in a graph, a path is a selection of features and the higher the centrality score, the most important the feature. It assigns a score of importance to each feature by taking into account all the possible feature subsets as paths on a graph. Correlation-based feature selection (CFS) sorts features according to pairwise correlations [186], whereas LASSO, proposed by Hagos et al. (2017), applies a regularization process that penalizes the coefficients of the regression variables while setting the less relevant ones to zero with respect to the constraint on the sum [187]. In LASSO, FS is a consequence of this process, when all the variables that still have non-zero coefficients are selected to be part of the model. Minimum redundancy maximum relevance (Mrmr) [188] is another well-known FS algorithm that systematically performs variable selection, achieving a reasonable trade-off between relevance and redundancy. A hybrid FS methodology was also employed that combines the outcomes of six FS techniques: two filter algorithms (Chi-square and Pearson correlation), three embedded ones (LightGBM, logistic regression and random forest) and one wrapper (with logistic regression) [163]. In this approach, all six FS techniques are applied separately, with each one resulting in a selected FS, and the final feature ranking is decided on the basis of a majority vote scheme. PCA is a well-known feature reduction method that reduces the

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

dimensionality of data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. The MATLAB-based feature selection library FSLib 2018 (https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library, accessed on 30 January 2021) was used for the implementation of the competing FS algorithms on a research workstation with Intel Core i7-7500 processor, 2.70 GHz CPU (16 GB RAM).

Figure 3.7 depicts the results of the comparison between the proposed GenWrapper FS and a classical wrapper FS technique. Specifically, the obtained mean 10FCV accuracies are shown with respect to the number of features as they have been ranked by the two compared approaches using two classifiers (LR and SVM). The following remarks can be extracted from Figure 3.7:

- GenWrapper significantly outperforms the classical wrapper FS, especially for a small number of selected features (up to 20). This superiority is proven for both SVM and LR;
- GenWrapper employing SVM gives the best overall performance (71.25% at 35 selected features).

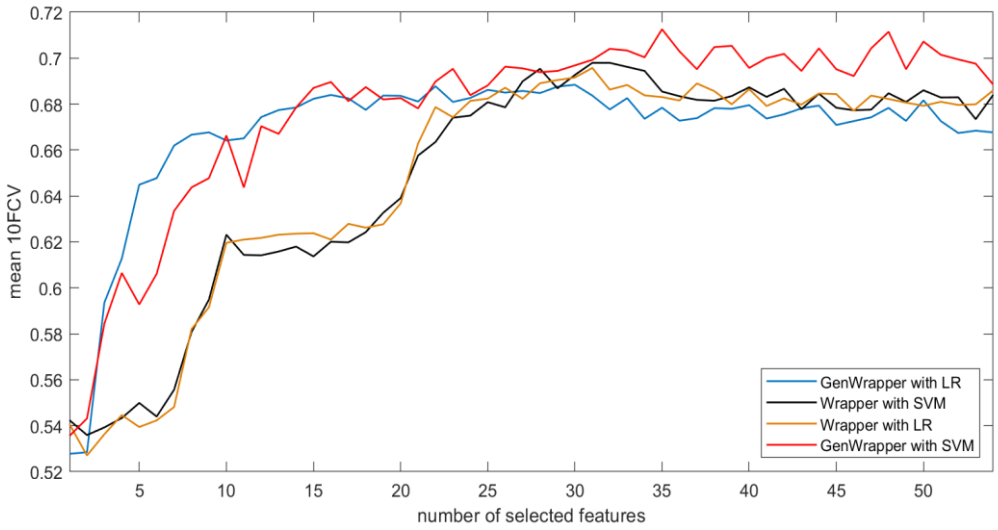


Figure 3.7. Accuracy (mean 10-fold cross-validation (10FCV)) with respect to selected features (curves): GenWrapper versus a classical wrapper using two classifiers (support vector machine (SVM) and logistic regression (LR)).

Figure 3.8 shows the progression of the mean 10FCV accuracy with respect to the number of selected features for the proposed FS and the other seven competing FS techniques (CFS, ILFS, Inf-FS, Lasso, Mrmr, PCA and hybrid). In this comparative

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

analysis, a linear SVM classifier were employed by all techniques since it proved to be the most efficient ML model. GenWrapper is the best-performing technique, achieving high accuracies (3.4% higher than the second best). Hybrid FS and Mrmr were the second and third best performers, achieving accuracies of 67.85% and 67.29%, respectively. Mrmr was very successful at the first 10 selected features but then it reached a threshold within the range of 67–68%, whereas the inclusion of further features had a minor or even negative effect on the classification performance. The rest of the FS techniques had moderate performances (61.97–65.11%). Table 3.5 also shows the best accuracies achieved by each technique and the number of features for which the best accuracy was achieved. GenWrapper achieved its best accuracy at a relatively small number of features (35), whereas the rest had inferior performances and, in most of the cases, at a higher number of features. The classical wrapper FS was the only one that selected slightly less features (31). A statistical comparison was finally conducted, verifying that the accuracies obtained by the proposed GenWrapper were significantly different (higher) to the ones of all the competing FS algorithms ($p < 0.001$).

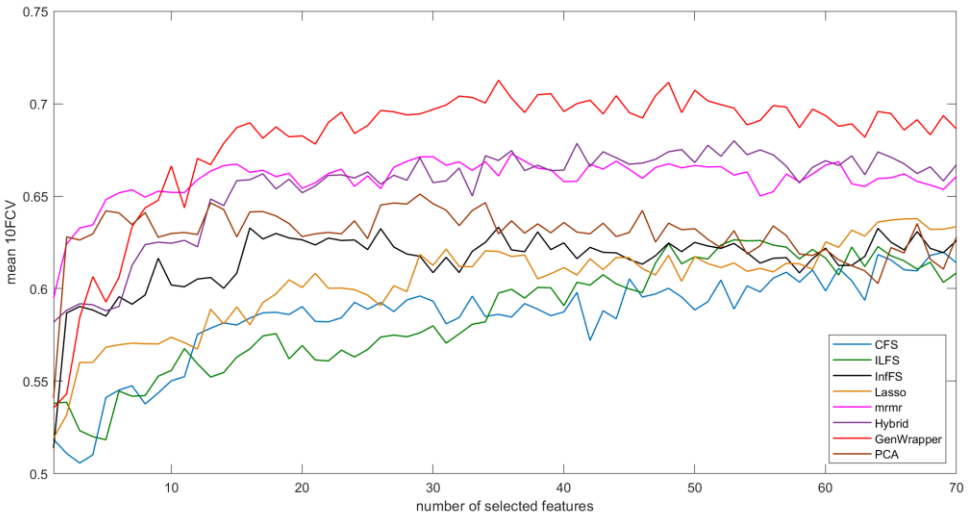


Figure 3.8. Accuracy (mean 10FCV) with respect to selected features: GenWrapper versus the remaining competing FS techniques. SVM was used for the classification task for all eight FS techniques.

Table 3.5. Best performance (mean 10FCV) achieved by all competing FS techniques employing SVM along with the number of selected features in which this accuracy was accomplished.

Approach	Best Accuracy (Mean 10FCV)	Number of Features	Statistical Comparison *	Execution Time (sec) **

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

GenWrapper	71.25	35	-	311.6
Wrapper	69.79	31	$p < 0.001$	10.2
CFS	61.97	69	$p < 0.001$	0.1
ILFS	63.63	82	$p < 0.001$	0.5
Inf-FS	63.32	35	$p < 0.001$	0.1
Lasso	64.41	94	$p < 0.001$	21.2
Mrmr	67.29	36	$p < 0.001$	2.3
Hybrid	67.85	41	$p < 0.001$	15.5
PCA	65.11	29	$p < 0.001$	<0.1

* Statistical comparison with the proposed GenWrapper. ** All the algorithms were executed on an Intel Core i7-7500 processor, 2.70 GHz CPU (16 GB RAM) using MATLAB 2020b.

The last part of the conducted comparative analysis focuses on a different performance metric—that is, the consistency of the obtained accuracies during the proposed repetitive validation process. As explained in the previous sections, the predictive capacity of the selected features is validated multiple times (10). In each of the ten repetitions, 10FCV is employed on a different, randomly selected balanced data sample. A feature subset could be considered as robust when it consistently leads to high accuracies over the ten repetitions. Figure 3.9 is a bar graph that visualizes (i) the mean 10FCV accuracies, (ii) the standard deviation of the 10FCV accuracies, (iii) the range ([min,max]) of the 10FCV accuracies and (iv) any outliers that deviate from the distribution of the 10FCV accuracies. GenWrapper was the most accurate approach (71.25%) and, at the same time, it proved to be the most consistent FS technique, with the great majority of obtained 10FCV accuracies being higher than 70%. The classical wrapper FS was also consistent over the ten repetitions but it was considerably less effective than the proposed GenWrapper. It should be noted that the hybrid FS approach achieved accuracies up to 72.5%; however, it does not generalize well given that it leads to a quite enlarged min–max range as well as an increased standard deviation, with the minimum accuracy being less than 60%. Mrmr has led to both moderate mean accuracy and moderate consistency (ranging between 66% and 70%) over the repetitions of the employed validation process. The rest of the competing FS approaches led to much lower 10FCV accuracies that ranged between 58% and 68%.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

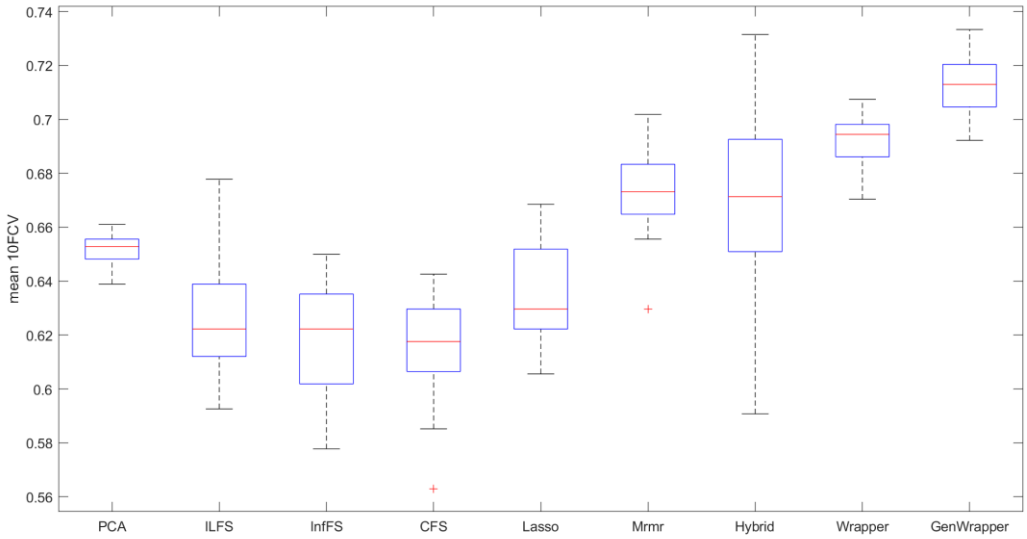
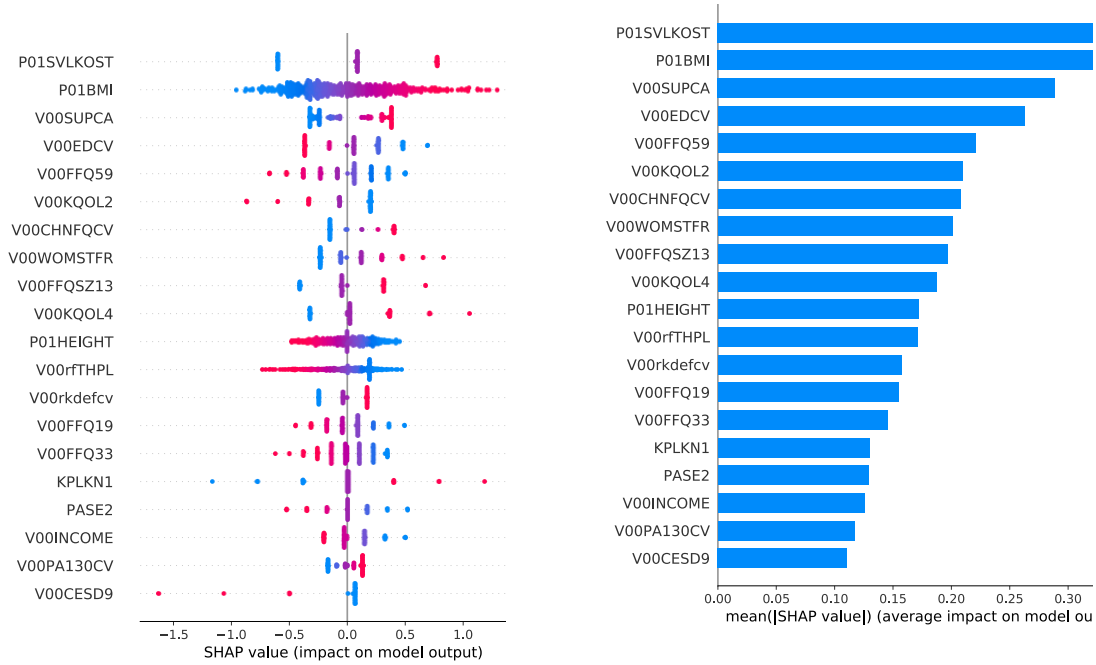


Figure 3.9. Bar graph comparison for the best models (SVMs trained on the optimum number of selected features per case). Red lines correspond to the mean 10FCV, blue boxes visualize the standard deviation of the obtained accuracies, dashed black lines show the min–max range and the red crosses depict outliers (if any).

Explainability Results

Figure 3.10a illustrates the features' impact on the output of the final model (SVM) on the OAI dataset. It sorts features by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the contribution of each feature (positive or negative) on the model's output. The color represents the feature value (blue—low; red—high). This reveals, for example, that a high P01BMI (body mass index of the participants) increases the predicted status of the participants. Similarly to BMI, the features P01SVLKOST, V00SUPCA, V00CHNFQCV, V00WOMSTFR, V00FFQSZ13, V00KQOL4, V00rkdefcv, KPLKN1 and V00PA130CV have a positive effect on the prediction outcome (their increase drives the output to increase), whereas the rest have the opposite effect. Figure 3.10b demonstrates the mean absolute value of the SHAP values which represents the SHAP global feature importance. It should be noted that the features P01SVLKOST, BMI, V00SUPCA and V00EDCV were the most important variables that significantly affected the prediction output (Appendix B).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186



(**a**) (**b**)

Figure 3.10. This figure depicts: (**a**) the SHAP summary plot and; (**b**) the SHAP feature importance for the SVM trained on the features selected by the proposed GenWrapper.

Discussion

Predicting KOA onset and its further progression is among the best strategies to reduce the burden of the disease. Risk factors for incident OA may differ from those for OA progression given that the incidence and progression of radiographic knee OA may involve different processes [189, 190]. Several risk factors have been reported to be associated with the incidence of knee OA [3, 191, 192]. However, our understanding about predictive risk factors associated with KOA progression is limited due to the fact that the number of studies, in which risk factors and incidence of knee OA have been investigated longitudinally, is relatively small. This study contributes to the identification of robust risk factors for knee OA progression as a first, but very important, step toward achieving the goal of developing preventive strategies and intervention programs and finally reducing the incidence and associated morbidity of knee OA.

Identifying important features from an imbalanced data set is an inherently challenging task, especially in the current KOA prediction problem with limited samples and a massive number of features. Feature selection algorithms employing

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

data resampling have been typically utilized to reduce the feature dimensionality and at the same time to overcome the class imbalance challenge. Oversampling algorithms randomly replicate examples from the minority class which in some scenarios can facilitate the FS process but is also prone to overfitting [193]. In data under-sampling, examples from the majority class are randomly discarded in order to rectify the disparities between classes. However, informative samples might be discarded from the final training set, reducing the generalization capabilities of the finally selected risk factors. New approaches are needed to address the intersection of the high dimensionality and imbalanced class problems due to their complicated interactions.

To cope with all the aforementioned challenges, the proposed FS technique incorporates a number of features aiming towards the identification of robust risk factors (with increased generalization capacity) extracted from a highly imbalanced dataset. GenWrapper relies on a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution. This stochastic search is employed to identify a globally optimal feature subset, compared to a costly search that makes local decisions. The proposed FS performs better than traditional feature selection techniques, can manage datasets with many features and does not need any specific knowledge about the problem under study. Compared to traditional GA-based FS algorithms, GenWrapper applies random undersampling at each individual solution, forcing the GA to converge to solutions (feature subsets) that generalize well regardless of the applied data sampling. K-fold cross-validation is utilized to measure the fitness of each individual solution, guaranteeing that the selected features have high predictive capacity over the whole dataset considered. Finally, instead of selecting the "best" individual of the final population, the proposed FS ranks features with respect to the number of times that they have been selected in all the individual solutions of the final population. This leads to selected features that consistently work well at any possible data sample and, thus, have increased generalization capacity with respect to KOA progression.

Linear classifiers were employed on this study, and this choice can be attributed to the fact that evidence of linear separability between the two classes (progressors versus non-progressors) was identified in previous studies of the authors on the same problem. Specifically, as it was reported in [163], LR and linear SVMs outperformed all the competing non-linear models (including Random Forest, XGboost, KNN and decision trees) on the same problem of predicting KOA. This finding highlight that the power of the proposed technique lies on the selection of robust and informative risk factors, whereas the complexity of the finally employed classification models plays a less crucial role.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

The performance of the proposed FS methodology was compared with eight well-known FS techniques in the recent literature. GenWrapper employing SVM led to the overall best performance (71.25% at 35 selected features), significantly outperforming all the competing algorithms. Specifically, it proved to be more accurate than the classical wrapper FS (which was the second-best approach), and this superiority was more evident for a small number of selected features (up to 20). GenWrapper was also much more effective (at least 3.4% more accurate) than the other seven competing FS techniques (CFS, ILFS, Inf-FS, LASSO, Mrmr, PCA and hybrid). Finally, apart from being the most accurate approach, GenWrapper was prove to also be the most consistent FS technique, with the great majority of the obtained 10FCV accuracies being higher than 70%, whereas all the other competing FS algorithms led to inferior and less consistent accuracies.

During our study, we utilized multimodal data and we managed to identify the variables that mainly contributed to the predictive ability of our models. Important predictive risk factors selected by our models included assessments of pain and function, qualitative assessments of X-rays, assessments of behavioral characteristics, medical history and nutrition from the Center for Epidemiologic Studies Depression Scale (CES-D) and Block Brief 2000 questionnaires. The strongest indicator variables are reporting on knee baseline radiographic OA status (P01SVLKOST), on anthropometric characteristics (P01BMI) and on nutritional (V00SUPCA) and behavioral habits (V00KQOL4). Previous studies [74, 79] have also reported similar key predicted variables for KOA progression. Our findings suggest that early functional, behavioral and nutritional interventions should be encouraged and implemented for the prevention or slowing-down of KOA progression.

Genetic algorithms might be costly in computational terms since the evaluation of each individual requires the training of a model. Due to its stochastic nature, the proposed FS takes a longer time to converge, and this could be considered as a limitation. However, the identification of risk factors for KOA progression is, in principle, an offline approach, and therefore, its current execution time (~5 min) is not prohibitive. In the current study, time execution is not considered as crucial as the predictive capability of the finally selected features that can be used to enhance our understanding of whether a patient is at increased risk of progressive KOA. GenWrapper improves the current state of the art by identifying risk factors that are more accurate compared to the ones selected by eight well-known FS algorithms (by at least 3.4%) and, most importantly, more robust in terms of their performance on the entire population of subjects (as it has been validated with an extensive validation mechanism that involved 100 training runs on different data samples). This stated improvement could (i) allow preventive actions to be planned and implemented and

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

(ii) enable more personalized treatment pathways and interventions for treatment, targeting specific risk factors. From a different perspective, being able to identify non-progressors could also prevent over-investigations and over-treatment.

Future work includes the identification of subpopulations of patients that have a greater risk of developing knee OA as well as a higher chance to progress faster. Moreover, quantification of KOA progression is another field that has not been adequately investigated by the scientific community. The combination of more advanced AI tools (e.g., Siamese neural networks) with the FS algorithm proposed in this study could form a reliable basis for quantifying KOA progression.

Conclusions

This study focuses on the identification of important and robust risk factors which contribute to KOA progression. The proposed FS methodology relies on an evolutionary machine learning methodology that leads to the selection of a relatively small feature subset (35 risk factors) which generalizes well on the whole dataset (mean accuracy of 71.25%). We investigated the effectiveness of the proposed approach in a comparative analysis with well-known FS techniques with respect to metrics related to both prediction accuracy and generalization capability. The nature of the selected features along with their impact on the prediction outcome (via SHAP) were also discussed to increase our understanding of their effect on KOA progression. Identifying and understanding the contribution of risk factors on KOA progression may enable the implementation of better prevention strategies prioritizing non-surgical treatments, essentially preventing an epidemic of KOA.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Funding

This research was funded by the European Community's H2020 Programme, under grant agreement Nr. 777159 (OACTIVE).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapter 4

Explainable Machine Learning for Knee Osteoarthritis Diagnosis Based on a Novel Fuzzy Feature Selection Methodology

Unpublished data:

This work has been submitted for publication.

Abstract

Knee Osteoarthritis (KOA) is a degenerative joint disease of the knee that results from the progressive loss of cartilage. Due to KOA's multifactorial nature and the poor understanding of its pathophysiology, there is a need for reliable tools that will reduce diagnostic errors made by clinicians. The existence of public databases has facilitated the advent of advanced analytics in KOA research however the heterogeneity of the available data along with the observed high feature dimensionality make this diagnosis task difficult. The objective of the present study is to provide a robust Feature Selection (FS) methodology that could: (i) handle the multidimensional nature of the available datasets and (ii) alleviate the defectiveness of existing feature selection techniques towards the identification of important risk factors which contribute to KOA diagnosis. For this aim, we used multidisciplinary data obtained from the Osteoarthritis Initiative database for individuals without or with KOA. The proposed fuzzy ensemble feature selection methodology aggregates the results of several FS algorithms (filter, wrapper and embedded ones) based on fuzzy logic. The effectiveness of the proposed methodology was evaluated using an extensive experimental setup that involved multiple competing FS algorithms and several well-known ML models. A 73.55 % classification accuracy was achieved by the best performing model (Random Forest classifier) on a group of twenty-one selected risk factors. Explainability analysis was finally performed to quantify the impact of the selected features on the model's output thus enhancing our understanding of the rationale behind the decision-making mechanism of the best model.

Keywords: KOA diagnosis; machine learning; clinical data; explainability; feature selection

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Introduction

Knee Osteoarthritis (KOA) is one of the most common types of osteoarthritis and musculoskeletal disorder. Being the 11th highest cause of disability globally, KOA is a multifactorial disease that results from mechanical and constitutional factors [194]. Obesity, age, gender, knee injuries and lifestyle are likely risk factors of KOA as they have been highlighted in the relevant recent literature [195]. In addition, swelling, pain and stiffness have been characterized as typical symptoms of the disease with irreversible cartilage damage being KOA's main consequence [3, 144, 159]. KOA is closely associated with a huge economic burden for the healthcare system and an unbearable health burden of the patients and their families. Significant consequences of KOA are the social isolation and low quality of life of the individual [160, 196]. Furthermore, the quantification of KOA is performed with the Kellgren–Lawrence (KL) severity grading scale, which is the most commonly grading system (current gold standard) and consists of five severity grades, from 0 to 4 [4].

Despite the fact that the scientific community has put a lot of effort into KOA research, a major challenge remains with respect to early diagnosis, long-term diagnosis and treatment of KOA. The parallel increase in computing power along with the collection of big datasets combined with the need to address the above challenges has led many research teams to use artificial intelligence (AI) techniques in the field of KOA [6]. In light of the above, several AI enabled studies have been proposed in the recent literature with the objective to diagnose or predict KOA. Yoo et al. used data from the Fifth Korea National Health and Nutrition Examination Surveys (KNHANES V-1) and the Osteoarthritis Initiative (OAI) to build an artificial neural network (ANN)-based a scoring system for the identification of KOA severity [95]. The proposed ANN model achieved an area under the curve (AUC) of 76% for the symptomatic KOA in an external validation with OAI data. In another study, Lim et al. proposed a method for early diagnosis of KOA based on clinical data from Korean National Health and Nutrition Examination Survey (KNHANES) [90]. They achieved a 76.8% AUC by using a deep neural network with scaled principal component analysis. In 2019, Christodoulou et al. investigated the deep learning capabilities in KOA diagnosis [197]. They used clinical data from OAI database and they achieved an 86.95% accuracy working on an aged subgroup (70 +).

In another study, Moustakidis et al. worked on self-reported clinical data (OAI) and proposed a deep learning methodology for the recognition of participants being at high risk of developing KOA in at least one knee and participants with symptomatic KOA [92]. They achieved accuracies up to 86.95%. Furthermore, Kwon et al. proposed an automatic classification of KOA severity that made use of gait analysis data and radiographic imaging (from Seoul National University Hospital) [198]. They

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

employed Inception-ResNet-v2 for feature extraction from X-rays and a support vector machine for KOA diagnosis achieving accuracies of 93%, 82%, 83%, 88% and 97% for the KL grades 0-4, respectively. In addition, Moustakidis et al. proposed a KOA classification approach with a focus on both accuracy and fairness [162]. They worked on different subgroups of participants from self-reported clinical data (OAI) and the dense neural networks methodology improved the accuracy up to 79.6% with fairness measured by balanced equalized odds (~ 92%) and demographic parity (98.5%) in the KOA case study.

Given that medical data and features can be subjective or difficult to interpret, medical decision making has a great potential to benefit from the use of fuzzy logic (FL). FL has been used to diagnose or facilitate decision making systems tackling many diseases, including OA. Hardi et al. proposed an expert system based on the fuzzy Tsukamoto method for OA diagnosis [199]. They treated symptoms of OA as fuzzy values that were further converted into firm value by using a weighted average demonstrating a 90% accuracy in the task of diagnosis of osteoarthritis disease. In general, various feature selection methods have integrated fuzzy logic in their internal mechanisms in order to handle the observed fuzziness and therefore improve the way that features are treated and combined. For instance, with emphasis to medical applications, the mutual information method combined with FL was used: (i) to select miRNAs in cancer [200]; (ii) to classify tumors [201]; and to select features for multilabel learning [202]. Similar studies include fuzzy entropy by using thresholds [203] for feature selection in various medical datasets and fuzzy rough sets [204, 205] for dimensionality reduction of feature space to prevent samples from misclassification.

It is well known that each one of the existing FS algorithms comes with its own advantages and disadvantages introducing a certain level of bias. To handle the multidimensional nature of the OAI dataset and to avoid bias and alleviate the defectiveness of single feature selection results, a fuzzy ensemble FS methodology is proposed in this work that aggregates the results of several FS algorithms (filter, wrapper and embedded). Fuzzy logic is employed to combine multiple feature importance scores thus leading to a more robust selection of informative features. The proposed method contributes to the significant reduction of the initial OAI feature dimensionality and to a decrease in the computational complexity of the classification models employed. To prove the effectiveness of the proposed methodology, an extensive experimental setup was designed involving multiple competing FS algorithms and several well-known ML models. As a post-hoc explainability, SHAP model was finally employed to identify the contribution of the selected features and the rationale behind the decision-making mechanism of best performing model.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Materials and Methods

Dataset Description

For the purpose of this study, data were obtained from the osteoarthritis initiative (OAI) database (available on https://nda.nih.gov/oai/). OAI is a prospective observational, multi-center and longitudinal study of KOA. OAI has enrolled 4796 women and men, aged 45-79 years. The present study used clinical evaluation data (643 features in total) from the baseline visit from all participants with or without KOA. The features of clinical dataset were divided into seven categories as shown in Table 4.1. Furthermore, in the present study, Kellgren and Lawrence (KL) grades were used as the outcome for the classification task.

Table 4.1. Main categories of the clinical evaluation data considered in this study.

Category	Description
Medical history	Medications and health histories based on questionnaire results (not included medical imaging outcomes)
Symptoms	Arthritis symptoms or health-related disability and function based on questionnaire data
Subject characteristics	Includes variables which describe anthropometric parameters and personal information
Nutrition	Questionnaire based on Block Food Frequency
Physical exam	Includes performance measures and knee and hand exams
Physical activity	Questionnaire results regarding living and leisure activities
Behavioral	Consists of variables which quantify the social behavior and the quality level of daily routine

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Methodology

The proposed AI methodology for KOA diagnosis consists of five processing steps: i) data pre-processing, ii) application of FS techniques, iii) learning process, iv) evaluation of the classification results and v) explainability analysis, as illustrated in Figure 4.1. An extensive explanation of the steps of the proposed methodology is given in the following subsections.

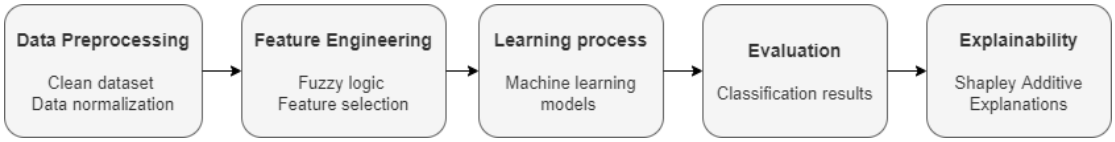


Figure 4.1. The proposed AI methodology for KOA diagnosis.

Problem Definition

In this study, we defined the KL-grade prediction task as a binary class classification problem. Specifically, the subjects of the study (3872 subjects in total) were divided into two equal groups:

i) KOA - participants who have KL >=2 at baseline. Participants in the group who had KL grades equal (early diagnosis) or higher than 2 in at least one of the two knees or in both at baseline;

ii) ii) non-KOA - participants who had KL0 or KL1 grade at baseline. Especially, this group of participants do not have KOA in any of their knees.

Data Pre-processing

Mode imputation was employed to handle categorical and continuous missing values [206]. In our study, data were normalised to [0, 1] to build a common basis for the FS algorithms and learning techniques that follow [207]. Furthermore, to cope with the imbalance data problem a stratified strategy for data resampling was applied. In particular, the number of the subjects in the majority class was reduced in order to become equal to the number of samples on the minority class [208].

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Proposed FS methodology

The proposed Fuzzy logic enhanced Feature Selection method (FLFS) combines the outputs of six well-known feature selection methods from three feature selection categories (Filter, Wrapper and Embedded). Specifically, from the filter category, the mutual information [209] and the f-ANOVA [210] techniques were applied. From the wrapper category, we employed a recursive feature elimination (RFE) based on logistic regression [211] and an RFE based on support vector machine [212] techniques, respectively. Furthermore, from the embedded category, a LightGBM [213] and a random forest technique [214] were applied. To calculate the importance of a feature for each category, the scores of the associated FS techniques were used as input to the Fuzzy Inference System (FIS) 1 that was implemented with Mamdani inference methodology [215]. The output of the FIS 1 was the defuzzification value that represents the feature importance score for the specific feature selection category. Then, the defuzzification score of each category was used as input to the FIS 2 where the output defuzzification value represents the overall feature importance. Figure 4.2 illustrates the FSFL flowchart with the defined fuzzy rules for each FIS and the selected feature selection methods for this study. Figure 4.3 shows the fuzzy sets used in the presented methodology for the input variables for FIS 1 and FIS 2, while Figure 4.4 shows the fuzzy sets of output variable for FIS 1 and 2.

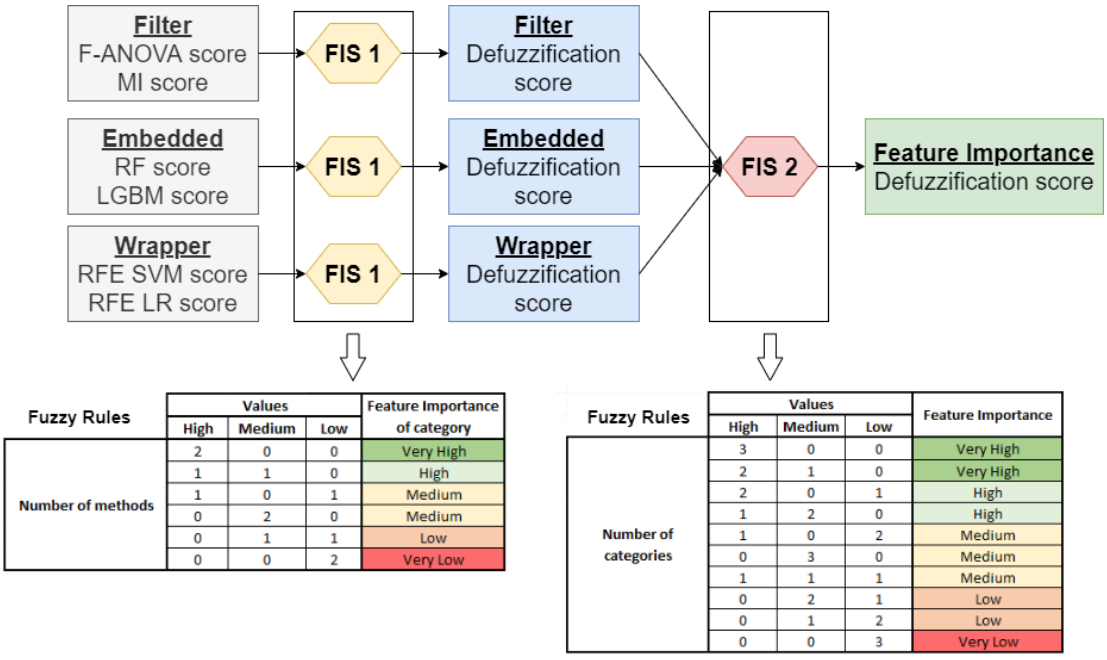


Figure 4.2. Feature Selection method based on Fuzzy Logic flowchart.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

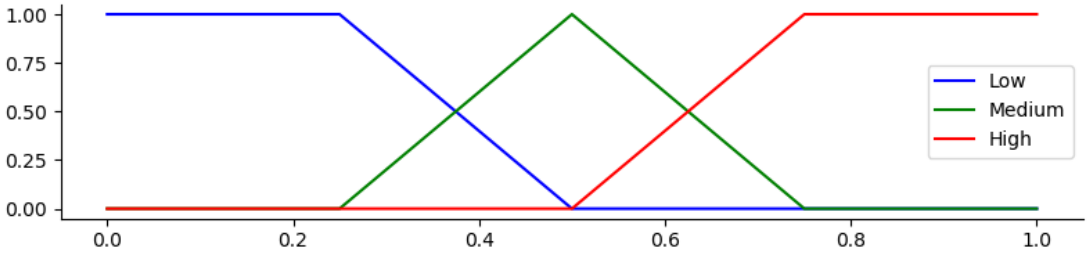


Figure 4.3. Fuzzy set of input variables for FIS 1 and 2.

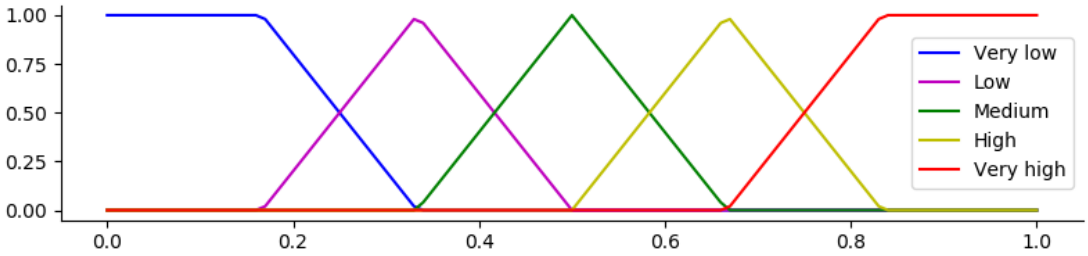


Figure 4.4. Fuzzy set of output variable for FIS 1 and 2.

Learning

In order to handle the demanding task of KOA classification, we investigated various ML models for their suitability and behavior in this problem. Specifically, random forest (RF) [216], multilayer perceptron (MLP) [217], logistic regression (LR) [161], support-vector machines (SVMs) [104], and k-nearest neighbors (KNN) [218] classifiers were tested. Furthermore, to avoid overfitting, and to optimize the performance of our models hyperparameter selection was applied individually per model.

Validation

For the experimental evaluation, a repeated stratified 5-fold cross validation was used [219]. The performance of the classifiers was also evaluated in terms of the recall, f1-score and precision as additional evaluation criteria [220]. A brief description of these metrics is given below. Initially, the accuracy is the ratio of correctly predicted observations to the total observations and can be characterized as the most intuitive performance measure. Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. Moreover, the ratio of correctly predicted positive observations to the total predicted positive observations is called precision or positive predictive value. F1-score is the weighted average of Precision and Recall.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Explainability

In the present work, we also examine how the risk factors have contributed to the final decision of KOA diagnosis. In order to achieve this, we used SHapley Additive exPlanations (SHAP), which is an approach to explain individual predictions based on Shapley Values of game theory and local explanations [183, 221]. In particular, we employed SHAP to rank features in terms of their impact on the final ML (Random Forest) outputs and to build a mini explainer model, which contributes to understanding the behavioral and the contribution of the risk factors in KOA diagnosis.

Results and Discussion

A. Results

In this section, we demonstrate the overall diagnosis performance of the models in relation to the first 100 selected features, and the highest metrics of the best models are also presented. Then, reference is made in the most important risk factors as they have been selected by the proposed Fuzzy FS methodology. Moreover, a comparative analysis is presented to prove the superiority of the proposed FS methodology compared to a number of well-known FS techniques. For the interpretation of the best model, an explainability analysis is employed to enhance our understanding of the reasoning behind its decision-making mechanism.

Diagnosis Performance

This subsection presents the results of a comparative analysis over a number of well-known ML models on the diagnosis classification task by using the first 100 selected risk factors. Figure 4.5 shows the testing accuracy performance (%) of the competing ML models with respect to the number of selected features. Specifically, KNN failed in diagnosis task, recording low testing accuracy performances. The rest of the ML models had an upward trend in the range of the first 15 risk factors. Overall, the best overall performance was achieved by RF with a maximum of 73.55% at 21 features.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

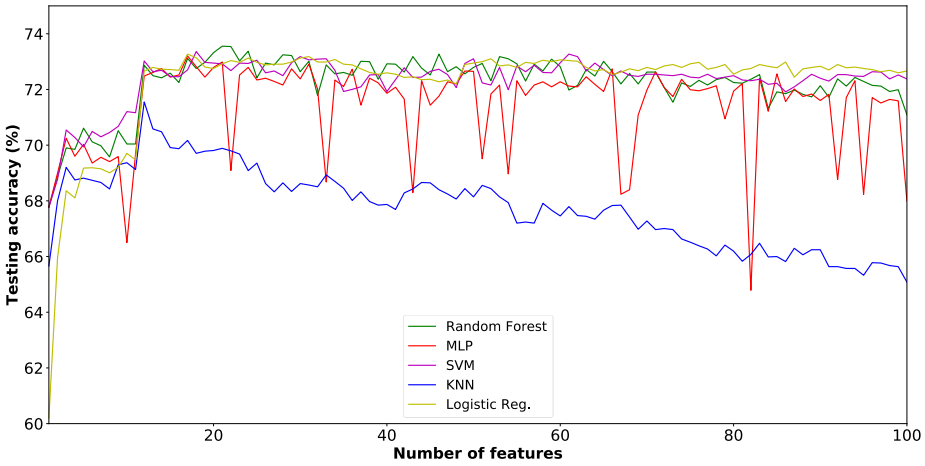


Figure 4.5. Curves with testing accuracy scores with respect to the number of selected features for different ML models.

Furthermore, the classification performance of the best performing ML models was further evaluated with respect to various validation metrics including class precision, recall, and f1-score. Table 4.2 demonstrates the best performance metrics of RF, MLP, LR, SVMs, and KNN models on the diagnosis task. In particular, RF achieved the best overall performance (73.55% accuracy) on the group of the twenty-one (21) risk factors. SVMs achieved the second-highest accuracy (73.36%). The rest of the ML models achieved lower accuracies.

Table 4.2. Summary of best metrics per model and number of selected features.

Models	Accuracy	Precision	Recall	F1-Score	Num. of Features
RF	73.55	73.82	73.64	73.59	21
MLP	73.20	73.48	73.20	73.13	17
LR	73.27	73.38	73.27	73.24	17
SVMs	73.36	73.68	73.36	73.27	18
KNN	71.55	71.74	71.55	71.49	12

Features Selected

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Figure 6 reveals more information about the origin of the 21 risk factors as selected by the chosen Fuzzy FS approach (Appendix C). As observed in Figure 4.6, six features describing subject characteristics were among the selected risk factors e.g., the age of the participants, the body mass index (BMI), and the diastolic blood pressure. Moreover, five out of the 21 selected risk factors come from the symptom's category, representing clinical parameters related to stiffness, knee difficulty, swelling, and pain, demonstrating the indication of the existence of KOA. Four of the risk factors are related to physical exams, whereas another two medical history and two physical activity parameters were selected as relevant to KOA occurrence. A behavioural risk factor and a nutrition risk factor were also selected by the proposed Fuzzy FS approach.

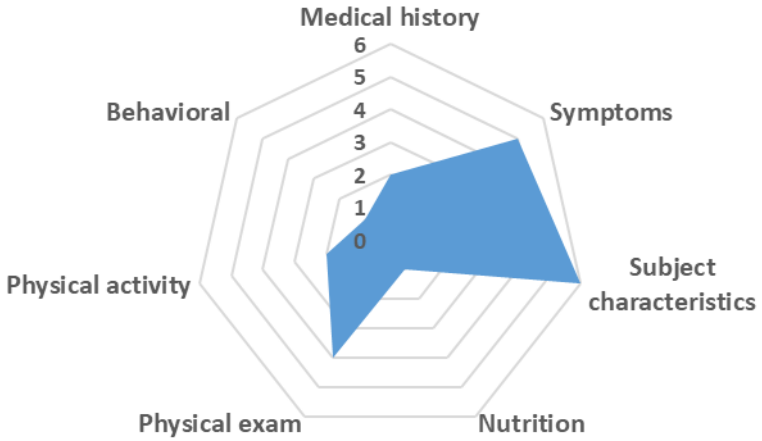


Figure 4.6. The 21 most informative selected risk factors per category.

Comparative Analysis

The performance of the proposed FSFL methodology was compared with each one of the six FS techniques that were also implemented independently. Finally, another recently published FS technique was also selected as comparative in which the final feature ranking, is decided on the basis of a majority vote scheme [163, 222].

Table 4.3 shows the maximum achieved accuracy in the first selected 100 features of OAI dataset and the number of features where the maximum accuracy was reached for each feature selection method used in the experimental evaluation with the best performed model (RF). The last row in Table 4.3 shows the dimensionality reduction achieved with the proposed FS method compared to other competitive methods. Specifically, the metric DR was defined to quantify the difference (%) in dimensionality reduction compared to FSFL:

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

$$DR = 1 - \frac{Max\ number\ of\ features\ (FSFL)}{Max\ number\ of\ features\ (Compared\ method)} \quad (1)$$

The proposed FSFL method achieved the best trade-off between performance and dimensionality reduction being capable of reducing significantly the feature set dimensionality while achieving slightly higher or comparable prediction performance with the rest of the competing algorithms. Specifically, the proposed FSFL technique reaches the highest accuracy (73.55%) at 21 selected features while the second-best accuracy (73.51%) was achieved by LBGM Emb at 87 features. This shows that the proposed FSFL technique results to a 76% smaller set of selected features compared to the second-best performing technique. On the other hand, the second-best performer with respect to dimensionality reduction was RF Emb with 73.36% accuracy achieved on a considerably larger feature subset with more than double features (43) compared to FSFL (21).

Table 4.3. Comparative analysis of FS methods.

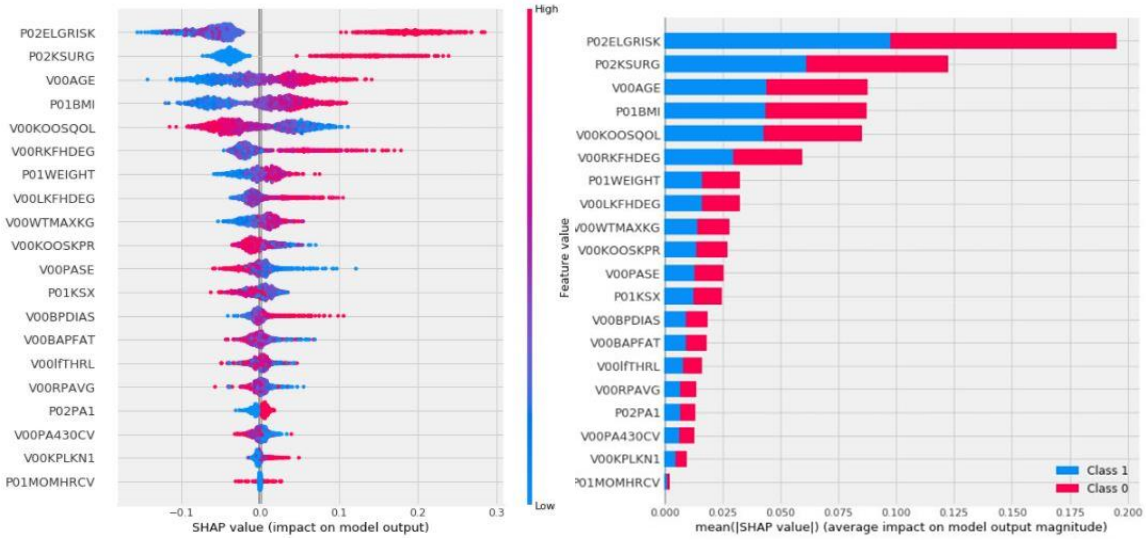
	FSFL	**Vote FS**	**RF Emb FS**	**LGBM Emb FS**	**SVM RFE FS**	**LR RFE FS**	**Filter MI FS**	**Filter f-ANOVA FS**
Maximum Accuracy (%)	73.55	72.99	73.36	73.51	70.53	73.50	72.75	73.44
Number of Selected Features	21	76	43	87	96	60	91	53
DR (%)	-	+72 %	+51%	+76%	+78%	+65%	+77%	+60%

Explainability Results

Figure 4.7a depicts how the features' impact shapes the output of the final model (RF) on the testing dataset. The features are sorted by the sum of SHAP value magnitudes over all testing subjects. Furthermore, the SHAP values are used to demonstrate the contribution of each risk factor (negative or positive) on the model's output. Specifically, blue color represents low feature values, whereas red color represents high values, respectively. In particular, a high value of PO2ELGRISK (knee symptoms, risk factors, or both status) increases the probability of the subjects to be assigned to

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

class KOA. Similarly to PO2ELGRISK, the higher the values of risk factors V00AGE, P02KSRG, P01BM1, V00RKFHDEG, P01WEIGHT, V00LKFHDEG, V00WTMACKG, V00BRDIAS, V00KPLKN1, and P02PA1, the more probable for subjects to belong to class KOA. The rest of the selected risk factors in Figure 4.7a have the opposite effect pushing the prediction output of the model to the class of healthy subjects. Figure 4.7b presents the SHAP global feature importance. The risk factors are sorted by the mean [|SHAP value|], which is the average impact on model output magnitude.



a) b)

Figure 4.7. a) Features' impact on Random Forest (21F) model output for the testing set of OAI dataset. b) Features' average impact magnitude for testing instances.

Figure 4.8 interprets locally the behavior of the model for the prediction output in a subject that suffers by KOA. P02ELGRISK (with a value of 2) and P01BMI (with a value of 29.8) push the predictions towards the class of KOA patients. Therefore, a high value of the aforementioned risk factors results to the increase of the output probability of the subject to be classified as KOA patient. On the contrary, increase of the risk factors P02KSURG, V00RKFHDEG, V00KOOSQOL, and V00KOOSKPR lowers the probability of a subject to be classified as KOA. Since, our prediction score = 0.51 > base value = 0.49, this subject has been positively classified, i.e., class KOA status.

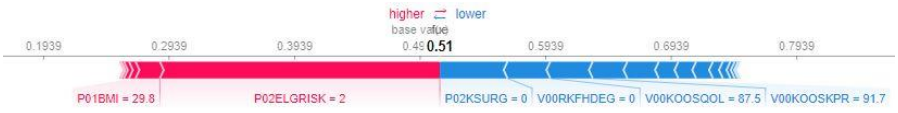


Figure 4.8. Risk factors contributions to ML model output for a KOA status subject

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

B. Discussion

Handling the multidimensional nature of the OAI dataset, a novel fuzzy ensemble FS methodology was designed, implemented and tested in this study. Its main novelty lies on the combination of several well-known FS algorithms based on a properly designed fuzzy inference mechanism that effectively aggregates their outputs. The superiority of the proposed FS technique was demonstrated through a thorough comparative investigation that included several state-of-the-art algorithms coming from different FS families (filter, wrapper, embedded and hybrid).

The proposed fuzzy FS methodology outperformed the aforementioned FS techniques achieving the best trade-off between dimensionality reduction and prediction accuracy. Working on a high-dimensional dataset of 643 features, twenty-one risk factors were selected for the objective of KOA diagnosis. Observing the nature of the selected risk factors, it was found that subject characteristics, symptoms, and physical exams are the most important risk factors contributing considerably to the KOA diagnosis. Overall, it was concluded that a combination of heterogeneous risk factors coming from different feature categories is needed for the effective diagnosis of KOA.

To sanity check the AI models beyond mere performance and further quantify the relevance of the selected risk factors, a post hoc explainability analysis was also conducted using SHAP. As observed by SHAP, P02ELGRISK, P02KSURG, V00AGE, P01BMI and V00KOOSQOL are five risk factors that have a major impact to the prediction output, which are in line with the existing literature. Specifically, P02ELGRISK, that represents knee symptoms, is an important risk factor in the diagnosis of KOA, as it has been identified by Lespasio et al. [2]. The history of knee surgery (P02KSURG) has been recognised as an important risk factor of KOA by Katz et al. [223], whereas the age of the subjects was also characterized as crucial in the occurrence of KOA and therefore was considered in the development of a predictive model for KOA diagnosis [92]. The knee injury and osteoarthritis outcome (KOOS) is a well-known knee-specific instrument that has been widely employed to evaluate quality of life in patients with knee injuries and identify patients who are at risk of developing OA [224]. Moreover, high BMI is suggested to be a high-risk factor in the development of KOA. High BMI values lead to the increment of knee joint mechanical loading [189].

Although the proposed FSFL technique selects a subset of risk factors with a significant dimensionality reduction compared to popular FS techniques, the application of a post-hoc explainability is still important in order to identify the contribution of the selected features to prediction output of the model. The use of explainability analysis algorithms for the interpretation of the ML models increases the understanding of the

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

principle of operation of each ML model and reveal the interactions that shape the diagnosis outcome.

The proposed methodology can be considered as computationally intensive; however, FS is considered here as an offline process and therefore the execution time does not play a crucial role. Future work will focus on the identification of easily measurable biomarkers and biomechanical parameters derived from musculoskeletal models, in combination with the already selected risk factors for the early diagnosis of KOA in the general population. Hence, to achieve this goal more advanced AI analytics tools in combination with the FSFL algorithm will be employed.

Conclusions

To enforce the development of more reliable, powerful, and non-invasive diagnostic tools, this study focuses on the identification and interpretation of the risk factors that contribute on the diagnosis of KOA. The proposed methodology is based on a novel fuzzy logic-based feature selection followed by learning algorithms and subsequently a post-hoc explainability analysis. The proposed technique aggregates the results of several FS algorithms (filter, wrapper and embedded ones), whereas fuzzy logic was employed to combine multiple feature importance scores thus leading to a more robust selection of informative features. The results showed that the presented methodology was capable to select a subset of risk factors that increase the performance accuracy of various ML models, compared to popular FS techniques. This was achieved with a significant decrease on the feature dimensionality (up to 78%). SHAP was finally applied to enhance our understanding of the rationale behind the decision-making mechanism of the selected ML model and the impact of the used risk factors on the prediction output.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Funding

This research was funded by the European Community's H2020 Programme, under grant agreement Nr. 777159 (OACTIVE).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapter 5

Leveraging explainable machine learning to identify gait biomechanical parameters associated with Anterior Cruciate Ligament injury

Unpublished data:

This work has been submitted for publication.

Abstract

Anterior cruciate ligament (ACL) tear is one of the most common knee injuries and it results in knee instability and increased risk of early onset osteoarthritis. ACL deficient and reconstructed knees display altered biomechanics during gait. Identifying significant gait changes is important for understanding normal and ACL function and is typically performed by statistical approaches. Unlike the existing techniques, this study focuses on the development of an explainable machine learning (ML) empowered methodology to: (i) identify important gait kinematic and kinetic parameters associated with ACL injury, (ii) quantify their contribution in the diagnosis of ACL injury and (iii) investigate the differences in sagittal plane kinematics and kinetics of the gait cycle between ACL deficient, ACL reconstructed and healthy individuals. For this aim, an extensive experimental setup was designed in which three-dimensional ground reaction forces and sagittal plane kinematic as well as kinetic parameters were collected from 151 subjects. The effectiveness of the proposed methodology was evaluated using a comparative analysis with seven well-known classifiers. A 94.95% classification accuracy was achieved by the best performing model (support vector machine) on a group of 21 selected biomechanical parameters. A state-of-the-art explainability analysis based on SHAP and conventional statistical analysis attempted to uncover the rationale behind the decision-making mechanism of the best trained model and provide a holistic approach of quantifying the contribution of the input biomechanical parameters in the diagnosis of ACL injury. Features, that would have been neglected by the traditional statistical analysis, were identified as contributing parameters having significant impact on the ML model's output for ACL injury during gait.

Keywords: ACL injury; walking biomechanics; machine learning; interpretation

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Introduction

Anterior cruciate ligament (ACL) tear is a frequent knee injury occurring in young active individuals during sport activities like basketball, football, ski and volleyball [225, 226]. The primary function of the ACL is to confine excessive posterior translation and external rotation of the femur relatively to the tibia against forces that act on the joint during gait and other activities [227-230]. As a result, an ACL deficient knee presents significant reflect on joint stability and biomechanics [231-233]. Studies utilising three-dimensional (3D) motion analysis have shown altered joint motion in ACL deficient knees during daily activities, such as walking, ascending and descending stairs or jumping [8, 234, 235]. This deviation causes a shift on the contact area and magnitude of shear forces at the knee joint which can lead to the initiation of osteoarthritis [236-239].

ACL reconstruction (ACLR) aims to lessen these changes in knee biomechanics. Annually 130.000 ACL reconstruction surgeries are performed in United States [240]. Although ACLR provides an improvement in knee stability and kinematics it is still questionable if the results are equal to pre-injury standards [241, 242]. As it was observed in several studies, increase or decrease in peak external knee-adduction moment, peak internal-rotation angle, increased medial contact force and decreased knee flexion angles were related to knee-joint cartilage loading and degeneration [236, 243-245]. Reductions in peak knee-flexion angle and external knee-flexion moment during the loading phase of gait have been reported at 6 to 60 months after ACLR [246-248].

Machine learning (ML) is an artificial intelligence (AI) analytic tool that constructs algorithms to identify patterns and characteristics contained within datasets. The goal is to train and validate prediction algorithms to achieve a desired result [249]. Musculoskeletal-specific models have already been developed to identify and classify fractures and predict functional outcomes after primary total knee arthroplasty (TKA) [250]. In 2017, Olczak et al. used deep learning techniques based on medical imaging to examine the feasibility of using AI to identify fractures in skeletal radiographs [251]. In another study, Kunze et al. based on partially modifiable risk factors developed ML algorithms to predict dissatisfaction after TKA [252]. Recent studies with individual-level datasets of gait analyses from kinetic skeletal tracking and advanced MR imaging (MRI) techniques focused on the determination of early progression of knee osteoarthritis (KOA) [253]. Moustakidis et al. proposed a novel fuzzy decision tree-based support vector machine (SVM) classifier by using 3-D ground reaction force (GRF) measurements to investigate KOA severity and to distinguish between asymptotic and osteoarthritis knee gait patterns [254]. Furthermore, Pedoia et al. performed ML multidimensional data analysis by using MR imaging and

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

biomechanical data [75]. They demonstrated that the analysis potentially indicates that cartilage composition may be an imaging biomarker for early KOA.

Machine learning approaches have been also used in studies to identify ACL injury based on MRI and biomechanical data or ACLR gait patterns with the aid of motion sensors. In 2017, Mazlan et al. proposed an ACL injury diagnosis system by using ACL injury MRI (normal, partial and crucial ACL) and SVM algorithm [255]. In another study, Chang et al. used MRI and deep learning techniques for the detection of complete ACL tear and achieved 96% test set accuracy [256]. Furthermore, Christian et al. used gait kinematics and ML techniques (SVM) to develop a pattern recognition system for diagnosis and evaluation of therapeutic treatment effect [257]. In another study, Zeng et al. proposed an approach for detection of the presence of ACL injury using kinematic features and neural networks [258]. Moreover, Todesco et al. proposed an ML approach for the identification of ACL gait patterns based on motion sensors data for on the field activities in rugby players [259].

Despite the relatively large number of ML studies on the field of ACL, the reported trained ML models are treated as black boxes. The lack of transparency and explainability of the models result to poor understanding of their inner workings and the rationale behind their decision-making mechanism. This work focuses on the development of an explainable ML-empowered methodology to identify important biomechanical parameters associated with ACL injury. The main contributions of this study are: (i) to examine how much each of the features contributed to the final ML decisions, (ii) to estimate the feature importance in the classification process and (iii) to investigate differences in sagittal plane kinematics and kinetics of the gait cycle between different patient groups based on a novel approach that combines explainable ML and statistical analytics. To achieve these goals, an extensive experimental setup was designed including biomechanical data collection, a thorough comparative analysis with seven well-known classifiers and a state-of-the-art explainability analysis.

Materials and Methods

Participants

A total of 151 subjects aged 18–50 years volunteered to participate in this study. Three different groups were defined: (a) ACL-deficient prior to surgery (ACLD), (b) ACL-reconstructed (ACLR) and (c) control (CON) group. All subjects were moderately active, participating in regular activity at least two times per week. The ACLD subjects

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

had suffered a unilateral ACL injury confirmed by an orthopedic surgeon and via magnetic resonance imaging. The ACLD group was examined an average of 30 days after injury, but before surgery. The ACLR subjects were included in the ACLR group if they had a unilateral ACL reconstruction and participated in the present study at least 6 months post-surgery. Individuals with different graft types (i.e., hamstring tendon and patellar tendon grafts) were included in the ACLR group. Both ACLD and ACLR subjects had a healthy contralateral knee, reported no other history of serious lower limb injury, and had resumed their physical activity at the time of the measurement. 53 subjects were recruited from the local community to serve as the CON group. The CON subjects were matched for age, gender, and physical activity status and had no history of ACL injury and neurologic disorder or other lower extremity injuries within 12 months prior to participating in the study. Prior to participation, all subjects signed a consent form, and all procedures were approved by the University of Thessaly ethics committee (approval code: 1660). The subjects' characteristics are presented in detail in Table 5.1.

Table 5.1. Subjects' characteristics.

Characteristics	ACLD	ACLR	CON
Gender	31 males and 13 females	40 males and 14 females	34 males and 19 females
Height	175.3±0.86 cm	177.6±0.80 cm	174.1±0.98 cm
Weight	77.38±14.91 kg	76.37±14.35 kg	72.23±15.81 kg

Testing procedure and data collection

Upon entering the gait laboratory, the subjects received instructions regarding the testing procedure and were familiarized with the walking task. ACLD and ACLR subjects completed the subjective Knee injury and Osteoarthritis Outcome Score (KOOS) evaluation form, which is considered a reliable measure of 5 outcomes, including activities of daily living, sport and recreation, pain, and knee-related quality of life [224]. Anthropometric measurements were recorded, and 20 spherical retroreflective markers were positioned bilaterally on anatomic landmarks and specific locations of the pelvis and lower limbs according to the marker set described in the literature [260, 261]. Subsequently, the subjects walked barefoot along the 10 m laboratory walkway within ±5% of their individual self-selected walking speed (SWS). SWS was measured during familiarization using infrared timing gates located in the middle of the walkway and was maintained throughout data collection via a

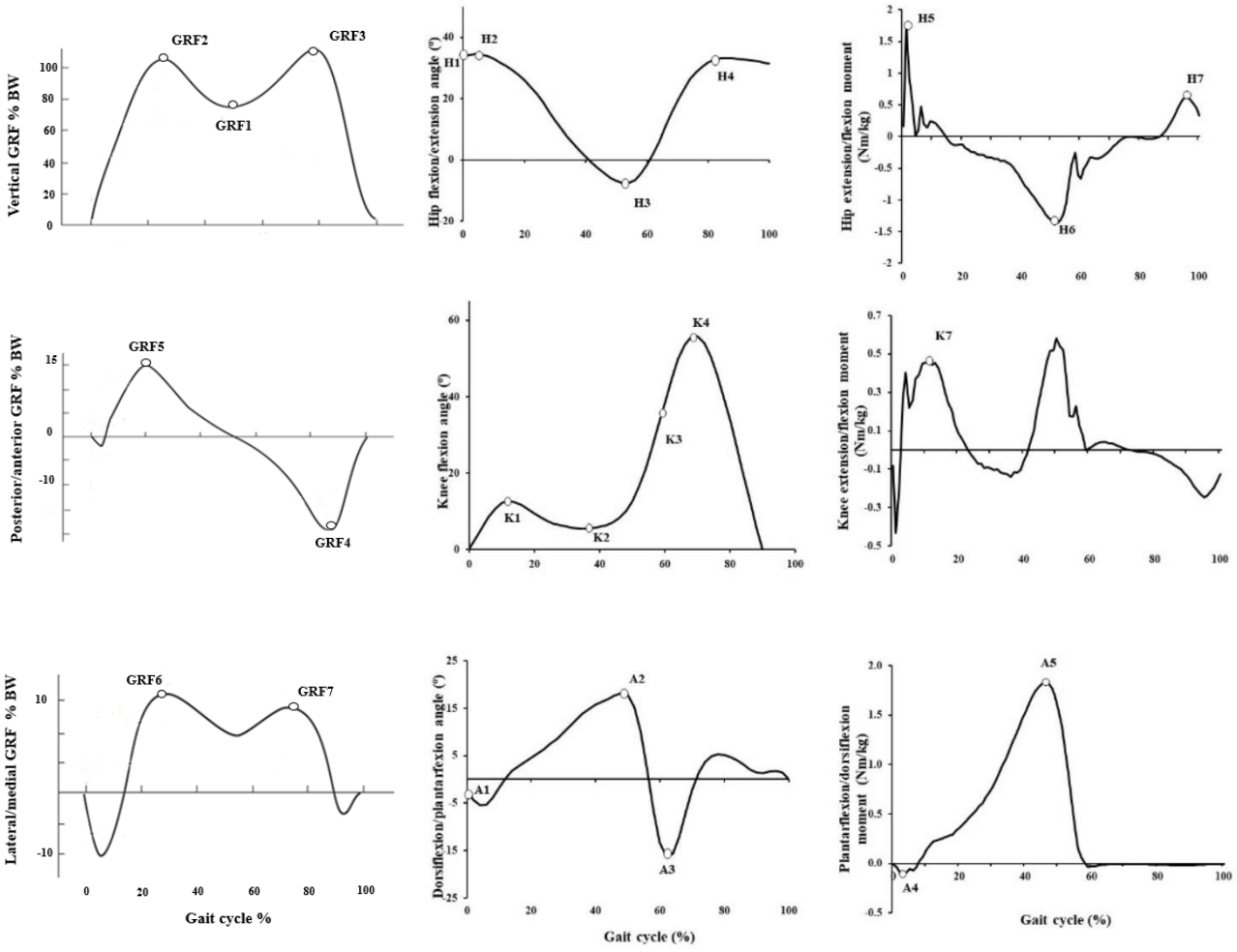
Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

metronome. Trials were performed until at least 5 complete gait cycles were recorded with each foot (left and right side) landing on the force platform. A trial was considered valid if the foot of the side being tested made a clean contact with the force platform located in the middle of the walkway and the walking speed was within ±5% of the individual SWS. Kinematic data were collected using 10 optoelectronic cameras (Vicon T-series, Oxford, UK) at 100 Hz and kinetic data were collected at 1000 Hz via a force platform (Bertec 4060-10, OH) embedded in the floor synchronized with the kinematic data.

Data Analysis

The symmetrical center of rotation estimation (SCoRE) [262] and the symmetrical axes of rotation approach (SARA) [263] were applied to optimize the calculation of the hip joint center and knee joint flexion axis, respectively. The initial contact and toe-off events of stance phase were determined from the vertical GRF (20N threshold) and the subsequent ipsilateral initial contact was determined from motion data using the Vicon Nexus software. Kinematic and GRF data were lowpass filtered with a 4th order Butterworth filter at 10 and 40 Hz, respectively. Inverse dynamics were used combining inertia properties of the segments as well as kinematic and GRF data to calculate net joint moments and powers of the lower limbs during the gait cycle. GRFs were expressed as a percentage of body weight, while net joint moments were expressed as internal moments and were normalized to body mass. Selected gait variables were extracted for each trial of each subject. A total of 155 trials were analysed for ACLD group, 204 trials for ACLR group and 298 trials for CON group, respectively. The three-dimensional GRFs, sagittal plane kinematic and kinetic variables of interest are presented in Figure 5.1 and Table 5.2. Data were analyzed from the ACLD/ACLR subjects' involved limb and for the control subjects, this was randomly assigned.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186



(a) (b) (c)

Figure 5.1. Three dimensional GRFs (a), sagittal plane kinematic (b) and kinetic (c) variables of interest during walking.

Table 5.2. Evaluated parameters of the gait cycle for vertical and horizontal GRFs and sagittal plane kinematics and kinetics.

Variables	Description
GRF1	Local minimum vertical GRF during support (% BW)
GRF2	First vertical GRF peak (% BW)
GRF3	Second vertical GRF peak (% BW)
GRF4	Anterior (propulsive) GRF peak (% BW)
GRF5	Posterior (braking) GRF peak (% BW)
GRF6	First lateral GRF peak (% BW)
GRF7	Second lateral GRF peak (% BW)
H1	Hip flexion angle at initial contact (°)
H2	Maximum hip flexion angle during stance phase (°)

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

H3	Maximum hip extension angle during stance phase (°)
H4	Maximum hip flexion angle during swing phase (°)
H5	Maximum hip extension moment during stance phase (Nm/kg)
H6	Maximum hip flexion moment during stance phase (Nm/kg)
H7	Maximum hip extension moment during swing phase (Nm/kg)
K1	Maximum knee flexion angle during stance phase (°)
K2	Minimum knee flexion angle during stance phase (°)
K3	Maximum knee flexion angle at foot off (°)
K4	Maximum knee flexion angle during swing phase (°)
K5	Knee flexion angle at first maximum knee extension moment during stance phase (°)
K6	Knee flexion angle at first vertical ground rection force peak (°)
K7	First maximum knee extension moment during stance phase (Nm/kg)
A1	Ankle angle at initial contact (°)
A2	Maximum dorsi-flexion angle during stance phase (°)
A3	Maximum plantar-flexion angle over the entire gait cycle (°)
A4	Maximum dorsiflexion moment during stance phase (Nm/kg)
A5	Maximum plantarflexion moment during stance phase (Nm/kg)

Machine Learning workflow

In order to identify knee kinematics associated with ACL injury, we designed, implemented and tested a multi-stage ML pipeline as shown in Figure 5.2. Its processing steps are presented as follows.

Figure 5.2. The proposed AI workflow for ACL diagnosis and interpretation.

Data were normalised to [0, 1] to build a common basis for the feature selection (FS) and the ML estimators. To rank our biomechanical parameters a well-established FS technique was applied. Relief algorithm [264] is a supervised learning algorithm and it is such effective in problems where strong dependencies between features are observed. Various well-known ML classifiers were evaluated for their suitability. Hyperparameter selection was implemented to avoid bias error, overfitting and optimize the performance of our ML models. Specifically, we used XGboost algorithm

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

[157] and Random Forest (RF) [265], which are ensemble learning algorithms and they are used due to their fast execution speed and increased model performance. Decision trees (DTs) were also evaluated providing a more interpretable decision-making mechanism [266]. Furthermore, we tested Naïve Bayes algorithm [267], which is based on applying Bayes' theorem and this method can be extremely fast. Being effective in high-dimensional spaces, SVM algorithms were also Included in our experimental analysis [268]. Moreover, Logistic Regression (LR) [269] and the K-Nearest Neighbor (KNN) algorithm [270] were tested. LR was employed to set the baseline performance obtained by a linear model and KNN was selected due to its ability to deal with the overfitting problems that appear in high-dimensional spaces.

For the evaluation of the proposed classifiers, a stochastic 70–30% random data split was applied to generate the training and testing subsets, respectively [163]. Specifically, the learning was performed on the stratified version of the training sets and the final performance was estimated on the accuracy testing sets. Furthermore, the performance of the classifiers was also evaluated in terms of the recall (or sensitivity), f1-score and precision as additional evaluation criteria [220].

In this work, we also: (i) investigated how much each of the features contributed to the final decision and (ii) estimated the feature importance. In order to achieve this, we used SHapley Additive exPlanations (SHAP) which are based on Shapley Values of game theory [183, 271]. SHAP offers the ability to interpret ML models, which are often treated as black boxes. In this study, we employed SHAP to rank features in terms of their impact on the final ML outputs and to build a mini explainer model. This enhances our understanding of the internal decision-making rationale of the trained AI models especially with respect to the mechanism with which selected biomechanical parameters are combined to produce decisions on ACL diagnosis and postoperatively.

Statistical Analysis

One-way analysis of variance (ANOVA) was used to investigate differences in sagittal plane kinematics and kinetics of gait cycle for the CON, ACLD and ACLR groups [272]. Furthermore, independent sample t-tests were employed to compare the first eight significant biomechanical parameters between the CON and the ACLD groups, which were indicated by the explainability analysis. On the same parameters, independent sample t-tests were also employed to evaluate the postoperative progress [273]. The significance level in our statistical comparisons was set at $p < 0.05$.

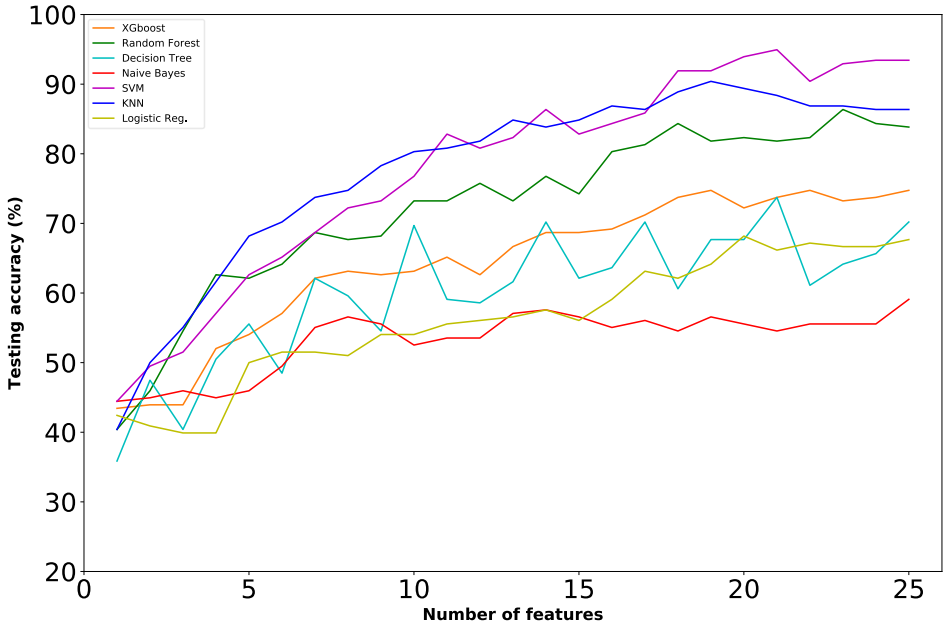
Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Results

Comparative Analysis

The proposed ML pipeline was initially applied on the three-class problem in which the patient groups CON, ACLD and ACLR are considered as separate classes. The proposed FS technique was executed on the pre-processed version of the 3-class dataset ranking the available features with respect to their relevance. The ML models were trained on feature subsets of increasing dimensionality (with a step of 1) and the testing classification accuracies were finally calculated until the full feature set has been tested. The classification results are given below.

Figure 5.3 demonstrates the accuracy testing performance (%) of the competing ML models with respect to the number of selected features on the 3-class problem. The majority of the ML models had an upward trend in the whole feature dimensionality range, followed by steady testing performance in most of the cases. Specifically, the SVM model showed an upward trend with respect to the first selected features, with a maximum of 94.95% (which was the overall best performance achieved). The second-best accuracy (90.40%) was achieved by the KNN model, which presented a steadily increasing performance. LR, DTs, RF and XGboost models also showed an upward trend with moderate accuracies ranging from 68.18% up to 74.5%. In contrast with the other models, Naïve Bayes failed in this task, recording low accuracy testing performances (in the range of 44.44–59.09%).



Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Figure 5.3. Learning curves with testing accuracy scores for different ML models trained on feature subsets of increasing dimensionality in the 3-class problem (refering to both ACL deficient and reconstructed patients).

Table 5.3 summarizes the results of XGboost, Random Forest, Decision Trees, Naive Bayes, SVM, KNN and Logistic regression on the three-class problem. The selected biomechanical parameters were in the range of 19-23 by the majority of the ML models (in six out of the seven), whereas the overall maximum was achieved by SVM on a group of twenty-one selected (21) biomechanical parameters. Naive Bayes selected less features (14) leading to low accuracy (57.58%). Furthermore, the second-highest accuracy was achieved by KNN (90.40%), whereas lower accuracies were obtained by RF, DTs and XGboost (less than 74.75%). Apart from being the most accurate overall, the SVM model recorded the best performance in the all metrics, namely precision (92.16%-96.72%), recall (92.19%-97.62%) and f1-score (93.07%-96.47%).

Table 5.3. Best testing accuracies (%) achieved for ML models in 3-class problem along with precision, recall, f1-score and the optimum number of features.

Models	Accuracy	Classes	Precision	Recall	F1-Score	Num. of Features
XGBoost	74.75	CON	70.80	95.24	81.22	19
		ACLD	81.48	44.00	57.14	
		ACLR	79.31	71.88	75.41	
Random Forest	86.36	CON	80.00	95.24	86.96	23
		ACLD	90.00	72.00	80.00	
		ACLR	94.83	85.94	90.16	
Decision Trees	73.74	CON	76.67	82.14	79.31	21
		ACLD	72.09	62.00	66.67	
		ACLR	70.77	71.88	71.32	
Naive Bayes	57.58	CON	65.69	79.76	72.04	14
		ACLD	40.91	54.00	46.55	
		ACLR	66.67	31.25	42.55	
SVM	94.95	CON	95.35	97.62	96.47	21
		ACLD	92.16	94.00	93.07	
		ACLR	96.72	92.19	94.40	
KNN	90.40	CON	85.26	96.43	90.50	19
		ACLD	95.12	78.00	85.71	
		ACLR	95.16	92.19	93.65	
Logistic Regression	68.18	CON	70.64	91.67	79.79	20
		ACLD	57.50	46.00	51.11	
		ACLR	71.43	54.69	61.95	

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Explainability Results

In this section, we interpret the contribution of the biomechanical parameters in shaping the AI model's output. To cope with this, we used explainability analysis on the best performing ML model (SVM). Initially, we performed a global investigation on the 3-class problem to quantify the overall features' contribution to the problem. Then, we performed explainability analysis on each one of the three trained binary (one-versus-one) SVM models that constitute the 3-class problem. Specifically, we applied SHAP analysis into the following three problems: i) control group versus ACLD group (local problem 1), ii) control group versus ACLR group (local problem 2), and iii) ACLD group versus ACLR group (local problem 3).

Global exploration

Figure 5.4 visualises the impact of the feature across all classes and the features were sorted by the sum of their SHAP values magnitudes across all instances. In this approach K2, H4, A3, GRF4, GRF7, K1, A4 and GRF6 were the parameters that affected the model output with mean SHAP values higher than 0.3.

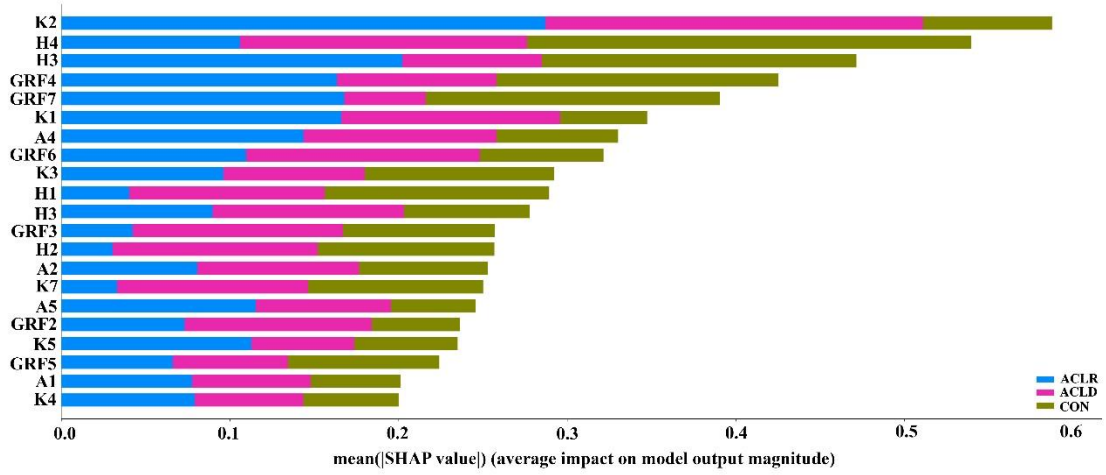


Figure 5.4. Average feature impact magnitude for all instances in the 3-class problem.

Local exploration

Figure 5.5 depicts the mean absolute value of the SHAP values which represents the SHAP global feature importance for local problem 1 (CON versus the ACLD). It should be noted that the features H4, K7, GRF3, H1, H2 were the most important

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

variables that significantly affected the prediction output. It is also observed that the contribution of H4 is 0.3 while the second-best parameter (K7) and all the remaining ones are below 0.18. From the above, H4 significantly contributes to the separation between the CON group and the ACLD group.

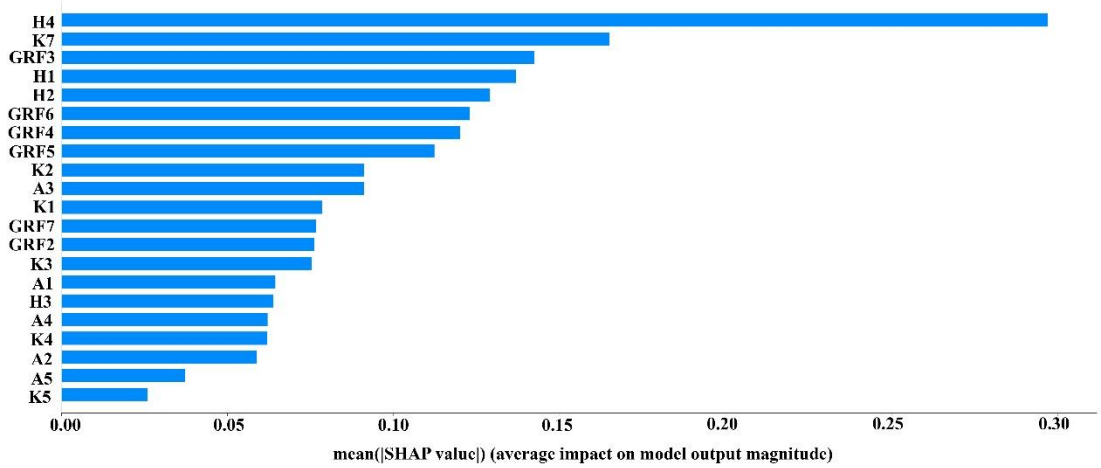
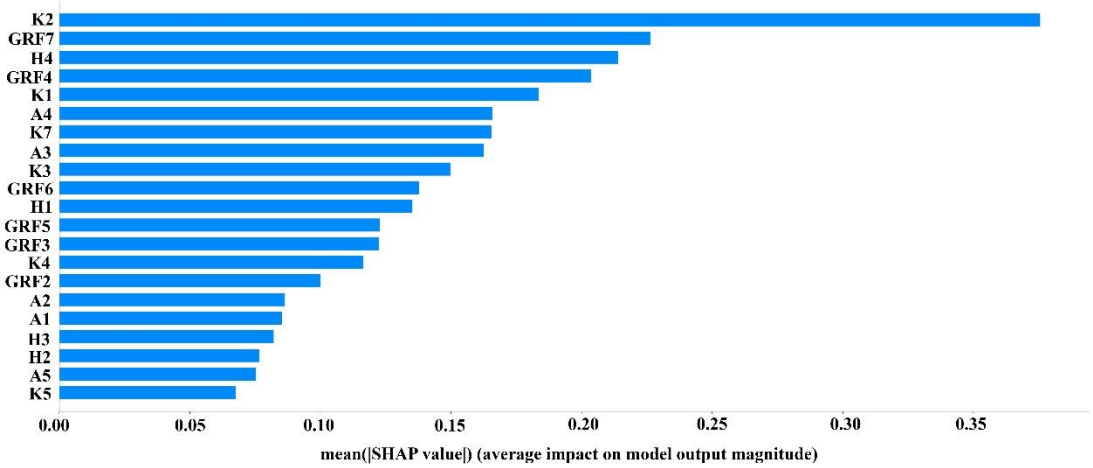


Figure 5.5. Features' impact on SVM model output for local problem 1. This figure shows the average impact magnitude for all instances in the task of differentiating the control group vs pre-surgery group.

Figure 5.6 depicts the mean absolute value of the SHAP values for local problem 2 that focuses on the discrimination of the CON and ACLR groups. Features K2, GRF7, H4, GRF4 and K1 were the most important variables that significantly affected the prediction output for the certain groups. Specifically, K2 records a much higher mean absolute value (higher than 0.35) compared to the rest of the features (that exhibit values less than 0.23).



Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Figure 5.6. Average feature impact magnitude for all instances in the local problem 2 (control versus ACLR)

The most important variables that significantly affected the prediction output in the local problem 3 (ACLD group versus ACLR group) were K2, H3, K7, A5 and A2, as shown in Figure 5.7. Similarly to local problem 2, parameter K2 is again the most important separation factor between individuals from the ACLD group and the ACLR group.

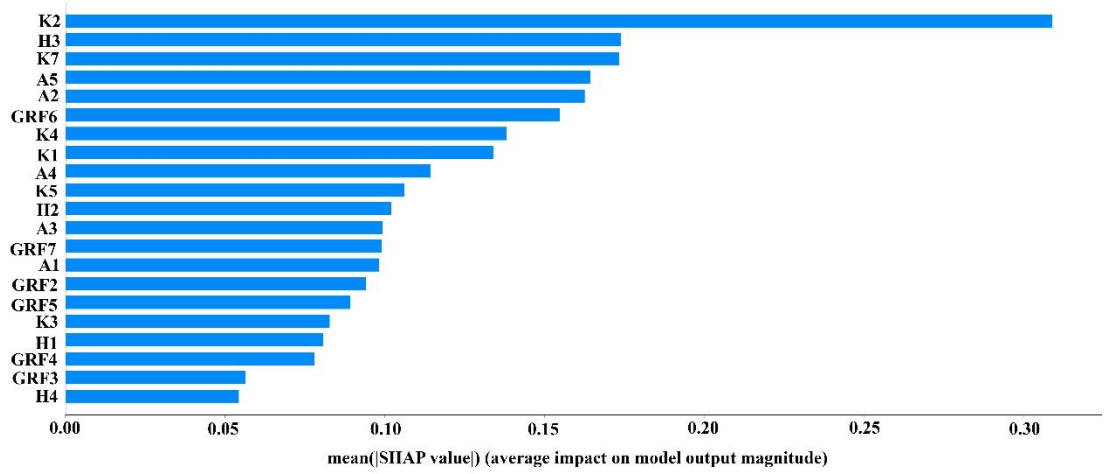


Figure 5.7. Average feature impact magnitude for all instances for local problem 3 (pre-surgery group versus post-surgery group).

Statistical Analysis

Statistical comparisons were also performed to identify whether there exist statistically significant differences between the classes considered for the most important features as they have been highlighted by the explainability analysis. Table 5.4 demonstrates the results of the one-way ANOVA tests, which performed to quantify these differences at the global level (with all three classes considered). As observed, there were statistically significant differences between the group means for all the comparisons.

Table 5.4. Statistical comparison at the global level.

		CON	ACLD	ACLR
Features	**Statistical Comparison**	**Mean (std)**	**Mean (std)**	**Mean (std)**

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

K2	$p < 0.05$	8.59 ± 3.81	8.35 ± 4.89	9.72 ± 4.70
H4	$p < 0.05$	37.93 ± 4.93	36.26 ± 6.42	37.03 ± 5.53
A3	$p < 0.05$	13.81 ± 6.52	15.91 ± 7.43	16.88 ± 7.25
GRF4	$p < 0.05$	19.61 ± 4.30	16.94 ± 5.17	16.76 ± 4.84
GRF7	$p < 0.05$	5.19 ± 1.65	5.81 ± 2.03	6.18 ± 2.92
K1	$p < 0.05$	21.62 ± 5.88	19.73 ± 6.56	19.88 ± 6.27
A4	$p < 0.05$	0.18 ± 0.08	0.19 ± 0.08	0.20 ± 0.09
GRF6	$p < 0.05$	5.69 ± 1.43	6.01 ± 2.21	6.36 ± 2.97

Then we performed statistical comparisons at the local level putting emphasis on the following tasks: (i) ACL diagnosis and (ii) rehabilitation after surgery. Initially, we run independent t-test analysis between the CON and the ACLD groups for the first eight significant biomechanical parameters, which were indicated by the explainability analysis of the specific binary problem (local problem 1). Subsequentially, we employed independent t-test analysis between the control and the ACL-reconstructed groups on the same parameters to identify which of them were modified and/or restored to their normal state (control level) as a measure of evaluating the postoperative progress.

Table 5.5 summarizes the results of the statistical analysis at the local level. The following remarks can be drawn from Table 5: (i) Significant differences were observed between CON and ACLD for half of the features considered, specifically the first three (H4, K7 and GRF3) along with GRF4; (ii) Four of the parameters (H1, H2, GRF6 and GRF5) that were considered important by the explainability analysis had no significant changes between CON and ACLD groups.

Table 5.5. Statistical analysis at the local level for ACL diagnosis and rehabilitation.

Features*	**CON vs ACLD**	**CON vs ACLR**
H4	$p < 0.05$	$p > 0.05$
K7	$p < 0.05$	$p < 0.05$
GRF3	$p < 0.05$	$p > 0.05$
H1	$p > 0.05$	$p > 0.05$
H2	$p > 0.05$	$p > 0.05$
GRF6	$p > 0.05$	$p < 0.05$
GRF4	$p < 0.05$	$p < 0.05$
GRF5	$p > 0.05$	$p > 0.05$

* Selected as important by the explainability analysis of the local problem 1 (CON versus ACLD)

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Discussion of Results

This work focuses on the development of a novel approach, which combines an explainable ML-empowered methodology and statistical analysis, for identifying important parameters associated with ACL injury. The problem has been coped as a three-class classification task where the participants of the study were divided into three groups (CON, ACLD and ACLR group). In addition to the classification part, the main contributions of this study are: (i) to investigate how much each of the features contributed to the final ML decisions, (ii) to estimate the feature importance in the classification process and (iii) to investigate differences in three dimensional GRFs, sagittal plane kinematics and kinetics of the gait cycle for the CON, ACLD and ACLR groups.

Being effective in problems with strong dependencies between features, the ReliefF algorithm was applied to serve as a FS technique and thus reduce the dimensionality of the initial feature space. Seven ML models were employed to perform the 3-class classification task on the reduced feature space where accuracies up to 94.95% were achieved. Specifically, the SVM model had the best performance and it showed an upward trend with respect to the first selected features, with a maximum of 94.95% at 21 features (which was the overall best performance achieved). Furthermore, the SVM model achieved rates from 92.16% up to 97.62% in each class for the metrics precision, recall and f1-score.

Having selected the most accurate ML model, this study attempted to uncover the rationale behind the decision-making mechanism of the trained model and therefore provide an alternative and a more holistic approach of quantifying the contribution of the input biomechanical parameters in the classification process. Specifically, explainability analysis was applied on the best performing ML model (SVM) and a global investigation was initially performed on the 3-class problem to quantify the overall features' contribution to the problem. As observed K2, H4, A3, GRF4, GRF7, K1, A4 and GRF6 were the most important biomechanical parameters that affected the model output. In order to estimate the feature importance separately, we also performed explainability analysis on each one of the three trained binary (one-versus-one) SVM models that constitute the 3-class problem. Specifically, we applied SHAP analysis into the following three problems: i) CON group versus ACLD group (local problem 1), ii) CON group versus ACLR group (local problem 2), and iii) ACLD group versus ACLR (local problem 3). As observed, in the local problem 1 the main biomechanical parameters were H4, K7 and GRF3. Furthermore, K2, GRF7 and H4 have the main contribution in local problem 2. In addition, from the third local problem K2, H3 and K7 have occurred as the most important biomechanical parameters. Previous studies have observed altered gait biomechanics in the ACL

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

deficient and ACL reconstructed patients compared to healthy individuals [8, 234]. These findings may indicate that the employed rehabilitation protocols fail to restore normal walking biomechanics, resulting in aberrant movement patterns. Several of the most important biomechanical parameters of ACL injury diagnosis highlighted by the global as well as the local explainability analysis used in our study coincide with the biomechanical outcomes reported in the literature to be related to altered gait patterns following ACLR. For example, maximum knee extension during stance phase (K7) significantly affected the prediction output in both of the aforementioned local problems examined in our study. K7 has been extensively investigated following ACLR and it has been consistently identified as a biomechanical parameter that is deceased following surgery and it is associated with poorer knee function in ACLR patients compared to healthy individuals [8, 274]. Additionally, minimum knee flexion angle during stance phase (K2) which had the most important contribution in local problem 2 and a significant one in local problem 3 has been reported to differentiate gait patterns between ACLR and healthy individuals up to 48 weeks post-surgery [274].

Besides explainability analysis, conventional statistical analysis was further performed to determine whether there exist significant differences between the three groups of our study for the aforementioned selected biomechanical parameters. As it was observed, in most of the cases the outcomes of the explainability and statistical analyses coincide. However, no significant differences were identified for many of those important parameters as shown in the case of local problem 1 (ACL diagnosis) in which H1, H2, GRF6 and GRF5 were identified as important by SHAP whereas their distributions had no significant differences between CON and ACLD. This finding implies that the proposed explainable ML methodology goes beyond the way that traditional statistics work. Features, that would have been neglected by the traditional statistical analysis, are highlighted as contributing parameters that have a significant impact on the ML model's output when they are combined with other statistically important ones. Moreover, as a measure of evaluating the postoperative progress, we performed statistical analysis for the local problem 1 and local problem 3 on the same parameters to identify which of them were modified and/or restored to their normal state (control level) after the surgery. Two of the three most important parameters (H4 and GRF3) were restored to their initial state after the surgery having no significant differences in the comparison between CON and ACLR groups. This means that these two biomechanical parameters (H4 and GRF3) were initially modified after the ACL injury and they were subsequently restored to their initial state after the surgery.

The clinical significance of our novel approach discussed in this work, which is based on a combination of an explainable ML-empowered methodology and statistical analysis to identify biomechanical parameters during walking associated with ACL

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

injury, should be considered with caution. This can be attributed to the fact that even though gait biomechanics are altered following ACLR, few biomechanical parameters demonstrate consistent results across studies and various tasks [234]. Factors such as, differences in the ACLR techniques (e.g. graft type), individual coping strategies among participants during walking, variations in employed rehabilitation protocols and gender differences may affect gait biomechanics alterations following ACLR as well as their clinical interpretation [8, 234, 275].

Explainability via SHAP or other similar tools is a crucial enabler allowing humans to better comprehend the decisions generated by black box models. However, SHAP is limited to simple explanations mainly quantifying the impact of individual features to the models' output [276]. Thus, the inner workings of the trained models and the way that the features are combined to reach the final decision remain hidden. Future work includes the combined use of graphical modelling with well-known explainability tools with the goal of identifying the relationships between features and the possible direct and indirect effect of features to the models' output. Such graphically-given explanations would enhance our understanding of the real rationale behind the decision-making mechanism of ML-empowered models acting on the tasks of ACL diagnosis and rehabilitation.

Conclusions

An explainable ML-empowered methodology was designed, implemented and tested in this study to identify important biomechanical parameters associated with ACL injury. The proposed extensive experimental setup included gait biomechanical data, a thorough comparative analysis with seven well-known classifiers and a state-of-the-art explainability analysis. According to the findings of the comparative analysis, a 94.95% classification accuracy was achieved by SVM on a group of twenty-one biomechanical parameters. The nature of the selected parameters along with their impact on the prediction outcome (via SHAP) were discussed to uncover the rationale behind the decision-making mechanism of the trained model and therefore provide an alternative and a more holistic approach of quantifying the contribution of the input parameters in the diagnosis of ACL injury. Statistical analysis was further performed to determine whether there exist significant differences between ACL deficient, ACL reconstructed and healthy individuals for the aforementioned parameters. Understanding the contribution of gait biomechanics is a valuable tool for creating more powerful and non-invasive prognostic tools in the hands of physicians, that will point out abnormal gait patterns in patients after ACLR to modify the rehabilitation protocol and avoid the development of osteoarthritis.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data available upon request.

Funding

This work was supported by the EC Horizon 2020 project OACTIVE Grant Agreement No. 777159, by the European Union and Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (SafeACL project, grant agreement T1EDK-04234) and by the Postgraduate Program of Study "Military Fitness & Wellbeing", School of Physical Education, Sports Science, University of Thessaly, Greece.

Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the University of Thessaly (protocol code 1660 and date of approval 03/06/2020).

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

General Conclusions

Our review outlined the current usage of machine learning methods in KOA diagnosis and prediction challenges. An increasing trend of ML-related studies and papers in the field of KOA indicate the need for (i) enhancing our understanding about the onset and progression of the disease and (ii) new data-driven tools that could enable early diagnosis and prediction of KOA. ML could play a key role towards these directions extracting valuable knowledge from various types of clinical data (biomechanical parameters, images, kinematics) and finding new solutions that utilize data from the greatest possible variety of sources. As far as the type of the ML models that were reported in our survey, SVMs were proved to be the most frequently used model in all the survey categories (21 Studies). The choice of SVM could be attributed to the fact that they generalize well in practice and that are computationally effective in high dimensional spaces. Neural networks were the second most frequent technique with three (3) studies reported for knee OA prediction and eighteen (18) applications of NN-based models in the OA classification survey. Machine learning can explore massive design spaces to identify correlations and multiscale modelling can predict system dynamics to identify causality. This has the potential to lead to the development of individually tailored treatments to maximize the efficacy of treatment.

The second chapter focused on the development of a ML-empowered methodology for KL grades prediction in healthy participants. The prediction task has been coped as a two-class classification problem where the participants of the study were divided into two groups (KOA progressors and non-progressors). Various ML models were employed to perform the binary classification task (KOA progressors versus non-progressors) where accuracies up to 74.07% were achieved. Moreover, we explored different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. Specifically, the overall best accuracy (74.07%) was obtained by combining datasets A and B that contain features from the baseline visit along with their progression over the next 12 months (Dataset D). Within the secondary objectives of this work were to identify informative risk factors from a big pool of available features that contribute more to the classification output (KOA prediction). As far as the nature of the selected features (55 risk factors), it was concluded that symptoms, medical imaging outcomes, nutrition and medical history are the most important risk factors contributing considerably to the KOA prediction. However, it was also extracted that a combination of heterogeneous features coming from almost all feature categories is needed to effectively predict KL progression.

In the third chapter we worked on a challenging task, to identify important risk factors which contribute to KOA progression from an imbalanced data set (OAI). Especially

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

in the current KOA prediction problem we used limited samples and a massive number of features. To cope with this aforementioned problem, we used data from the baseline visit along with progression data within the first 12 months (Dataset D, from Chapter 2) and we proposed an evolutionary machine learning methodology (GA-based wrapper technique) that led to the selection of a relatively small feature subset (35 risk factors) which generalizes well on the whole dataset (mean accuracy of 71.25%). Furthermore, the nature of the selected features along with their impact on the prediction outcome (via SHAP) were also discussed to increase our understanding of their effect on KOA progression. So, our findings suggest that early functional, behavioral and nutritional interventions should be encouraged and implemented for the prevention or slowing-down of KOA progression. Specifically, important predictive risk factors selected by our models are the following: assessments of pain and function, qualitative assessments of X-rays, assessments of behavioral characteristics, medical history and nutrition from the Center for Epidemiologic Studies Depression Scale (CES-D) and Block Brief 2000 questionnaires. The strongest indicator variables are the following: knee baseline radiographic OA status (P01SVLKOST), anthropometric characteristics (P01BMI) and nutritional (V00SUPCA) and behavioral habits (V00KQOL4). Previous studies [74, 79] have also reported similar key predicted variables for KOA progression.

The fourth chapter concerns the diagnosis task. The heterogeneity of the available bid data (OAI database) along with the observed high feature dimensionality make this diagnosis task difficult. To cope and to enforce the development of more reliable and non-invasive diagnostic tools, we worked on the identification and interpretation of the risk factors that contribute on the diagnosis of KOA. So, we proposed a methodology, which is based on a novel fuzzy logic-based feature selection followed by learning algorithms and subsequently a post-hoc explainability analysis. With respect to the nature of the selected risk factors, it was concluded that subject characteristics, symptoms, and physical exams are the most important risk factors contributing considerably to the KOA diagnosis. In order to sanity check the AI models beyond mere performance and further quantify the relevance of the selected risk factors, a post hoc explainability analysis was also conducted using SHAP. As observed by SHAP, P02ELGRISK, P02KSURG, V00AGE, P01BMI and V00KOOSQOL are five risk factors that have a major impact to the prediction output, which are in line with the existing literature.

The abnormal knee kinematics and kinetics after ACLR contribute to degenerative processes and they are characterized as risk factors for the progression of KOA. In Chapter 5 we developed a novel approach, which combines an explainable ML-empowered methodology and statistical analysis, for identifying important parameters associated with ACL diagnosis and postoperatively. A 94.95%

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

classification accuracy was achieved by the best performing model (support vector machine) on a group of 21 selected biomechanical parameters. A state-of-the-art explainability analysis based on SHAP and conventional statistical analysis attempted to uncover the rationale behind the decision-making mechanism of the best trained model and provide a holistic approach of quantifying the contribution of the input biomechanical parameters in the certain tasks. Several of the most important biomechanical parameters of ACL diagnosis and postoperatively highlighted by the global as well as the local explainability analysis used in our study coincide with the biomechanical outcomes reported in the literature to be related to altered gait patterns following ACLR. Despite the fact that parameters as H1, H2, GRF6 and GRF5 were identified as important by SHAP had no significant statistical differences. This finding implies that the proposed explainable ML methodology goes beyond the way that traditional statistics work. So, features that would have been neglected by the traditional statistical analysis, were identified as contributing parameters having significant impact on the ML model's output for ACL diagnosis and postoperativelyduring gait.

As a result, research work at the intersection of machine learning and KOA offers great promise for improving clinical decision-making and accelerating relevant intervention programs. For our future work, we are planning to also consider image-based biomarkers and areas with valuable information derived from biomechanical data that are expected to further improve the predictive capacity of the proposed methodology for the KL grades prediction. Furthermore, we are planning to work on the identification of subpopulations of patients that have a greater risk of developing knee OA as well as a higher chance to progress faster. The combination of more advanced AI tools (e.g., Siamese neural networks) with the proposed GA-based wrapper technique algorithm could form a reliable basis for quantifying KOA progression. In addition, for the diagnosis task future work will focus on the identification of easily measurable biomarkers and biomechanical parameters derived from musculoskeletal models, in combination with the already selected risk factors for the early diagnosis of KOA in the general population. Hence, to achieve this goal more advanced AI analytics tools in combination with the FSFL algorithm will be employed. At last, but not least in the task of ACL diagnosis and post-surgical rehabilitation, future work includes the combined use of graphical modelling with well-known explainability tools with the goal of identifying the relationships between features and the possible direct and indirect effect of features to the models' output. Such graphically-given explanations would enhance our understanding of the real rationale behind the decision-making mechanism of ML-empowered models acting on the tasks of ACL diagnosis and rehabilitation. The significant biomechanical parameters that will emerge will be an entry into the development of robust predictive models for the outset of the KOA.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

References

1. Cieza, A., et al., *Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10267): p. 2006-2017.
2. Lespasio, M.J., et al., *Knee osteoarthritis: a primer.* The Permanente Journal, 2017. **21**.
3. Silverwood, V., et al., *Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis.* Osteoarthritis and cartilage, 2015. **23**(4): p. 507-515.
4. Kohn, M.D., A.A. Sassoon, and N.D. Fernando, *Classifications in brief: Kellgren-Lawrence classification of osteoarthritis.* Clinical Orthopaedics and Related Research®, 2016. **474**(8): p. 1886-1893.
5. Jamshidi, A., J.-P. Pelletier, and J. Martel-Pelletier, *Machine-learning-based patient-specific prediction models for knee osteoarthritis.* Nature Reviews Rheumatology, 2019. **15**(1): p. 49-60.
6. Kokkotis, C., et al., *Machine Learning in Knee Osteoarthritis: A Review.* Osteoarthritis and Cartilage Open, 2020: p. 100069.
7. Jang, S., K. Lee, and J.H. Ju, *Recent Updates of Diagnosis, Pathophysiology, and Treatment on Osteoarthritis of the Knee.* International Journal of Molecular Sciences, 2021. **22**(5): p. 2619.
8. Hart, H.F., et al., *Knee kinematics and joint moments during gait following anterior cruciate ligament reconstruction: a systematic review and meta-analysis.* Br J Sports Med, 2016. **50**(10): p. 597-612.
9. Courties, A., J. Sellam, and F. Berenbaum, *Metabolic syndrome-associated osteoarthritis.* Current opinion in rheumatology, 2017. **29**(2): p. 214-222.
10. Cabitza, F., A. Locoro, and G. Banfi, *Machine learning in orthopedics: a literature review.* Frontiers in bioengineering and biotechnology, 2018. **6**.
11. Staugaard, A.C., *Robotics and AI: an introduction to applied machine intelligence*. 1987: Prentice-Hall Englewood Cliffs.
12. Ye, Q.-H., et al., *Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning.* Nature medicine, 2003. **9**(4): p. 416.
13. Helma, C., et al., *Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds.* Journal of chemical information and computer sciences, 2004. **44**(4): p. 1402-1411.
14. Larranaga, P., et al., *Machine learning in bioinformatics.* Briefings in bioinformatics, 2006. **7**(1): p. 86-112.
15. Voyant, C., et al., *Machine learning methods for solar radiation forecasting: A review.* Renewable Energy, 2017. **105**: p. 569-582.
16. Behmann, J., et al., *A review of advanced machine learning methods for the detection of biotic stress in precision crop protection.* Precision Agriculture, 2015. **16**(3): p. 239-260.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

17. Mullainathan, S. and J. Spiess, *Machine learning: an applied econometric approach.* Journal of Economic Perspectives, 2017. **31**(2): p. 87-106.
18. Kluzek, S. and T.A. Mattei, *Machine-learning for osteoarthritis research.* Osteoarthritis and cartilage, 2019. **27**(7): p. 977-978.
19. Zheng, A. and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. 2018: " O'Reilly Media, Inc.".
20. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
21. Dayan, P., M. Sahani, and G. Deback, *Unsupervised learning.* The MIT encyclopedia of the cognitive sciences, 1999: p. 857-859.
22. Noroozi, M. and P. Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*. 2016. Cham: Springer International Publishing.
23. Cox, D.R., *The Regression Analysis of Binary Sequences.* J. R. Stat. Soc. Ser. B 1958. **20**: p. 215-242.
24. Efroymson, M.A., *Multiple regression analysis. Math. methods Digit.* Comput, 1960. **1**: p. 191-203.
25. Craven, B.D.I., S. M. N. , *Ordinary least-squares regression.* SAGE Dict. Quant. Manag. Res., 2011: p. 224-228.
26. Friedman, J.H., *Multivariate Adaptive Regression Splines.* Ann. Stat. , 1991. **19**: p. 1-67.
27. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**(368): p. 829-836.
28. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901. **2**(11): p. 559-572.
29. Wold, H., *Partial Least Squares*. Vol. 6. 1985, 581–591 In Encyclopedia of Statistical Sciences.
30. Fisher, R.A., *The use of multiple measures in taxonomic problems.* Ann. Eugen, 1936. **7**: p. 179–188.
31. Tryon, R.C., *Communality of a variable: Formulation by cluster analysis.* Psychometrika, 1957. **22**: p. 241–260.
32. P. Lloyd, S., *Least Squares Quantization in PCM's*. Vol. 28. 1982. 129-136.
33. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-254.
34. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1-38.
35. Russell, S.J.N., P. , *Artificial Intelligence: A Modern Approach*. Vol. 9. 1995.
36. Duda, R.O., *Pattern classification and scene analysis / Richard O. Duda, Peter E. Hart*, ed. P.E. Hart. 1973, New York: Wiley.
37. Neapolitan, R.E., *Models for reasoning under uncertainty.* Appl. Artif. Intell., 1987. **1**(4): p. 337-366.
38. Breiman, L., *Bagging predictors.* Machine Learning, 1996. **24**(2): p. 123-140.
39. Schapire, R.E., *A brief introduction to boosting*, in *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*. 1999, Morgan Kaufmann Publishers Inc.: Stockholm, Sweden. p. 1401-1406.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

40. Freund, Y.S., R. R. E. , *Experiments with a New Boosting Algorithm*, in *International Conference on Machine Learning*. 1996. p. 148–156.
41. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
42. Belson, W.A., *Matching and Prediction on the Principle of Biological Classification.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1959. **8**(2): p. 65-75.
43. Breiman, L.F., J. H.; Olshen, R. A.; Stone, C. J., *Classification and Regression Trees*. Vol. 19. 1984.
44. Kass, G.V., *An Exploratory Technique for Investigating Large Quantities of Categorical Data.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1980. **29**(2): p. 119-127.
45. Friedman, J.H., *Stochastic gradient boosting.* Comput. Stat. Data Anal., 2002. **38**(4): p. 367-378.
46. Broomhead, D.S. and D. Lowe, *Multivariable Functional Interpolation and Adaptive Networks.* Complex Systems 2, 1988: p. 321-355.
47. Rosenblatt, F., *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol Rev, 1958. **65**(6): p. 386-408.
48. Linnainmaa, S., *Taylor expansion of the accumulated rounding error.* BIT Numerical Mathematics, 1976. **16**(2): p. 146-160.
49. Riedmiller, M. and H. Braun. *A direct adaptive method for faster backpropagation learning: the RPROP algorithm*. in *IEEE International Conference on Neural Networks*. 1993.
50. Hecht-Nielsen, R., *Counterpropagation networks.* Appl Opt, 1987. **26**(23): p. 4979-83.
51. Jang, J.R., *ANFIS: adaptive-network-based fuzzy inference system.* IEEE Transactions on Systems, Man, and Cybernetics, 1993. **23**(3): p. 665-685.
52. Melssen, W., R. Wehrens, and L. Buydens, *Supervised Kohonen networks for classification problems.* Chemometrics and Intelligent Laboratory Systems, 2006. **83**(2): p. 99-113.
53. Hopfield, J.J., *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the National Academy of Sciences, 1982. **79**(8): p. 2554-2558.
54. Pal, S.K. and S. Mitra, *Multilayer perceptron, fuzzy sets, and classification.* IEEE Transactions on Neural Networks, 1992. **3**(5): p. 683-697.
55. Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew, *Extreme learning machine: Theory and applications.* Neurocomputing, 2006. **70**(1): p. 489-501.
56. Goodfellow, I.B., Y.; Courville, A. , *Regularization for Deep Learning*. Deep Learning. 2016.
57. Salakhutdinov, R. and G. Hinton, *Deep Boltzmann Machines*, in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, D. David van and W. Max, Editors. 2009, PMLR: Proceedings of Machine Learning Research. p. 448--455.
58. Vincent, P., et al., *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.* Journal of machine learning research, 2010. **11**(12).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

59. Fix, E. and J.L. Hodges, *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.* International Statistical Review / Revue Internationale de Statistique, 1989. **57**(3): p. 238-247.
60. Atkeson, C.G., A.W. Moore, and S. Schaal, *Locally Weighted Learning.* Artificial Intelligence Review, 1997. **11**(1): p. 11-73.
61. Kohonen, T., *Learning vector quantization.* Neural Networks 1988. **1**: p. 303.
62. Kohonen, T., *The self-organizing map.* Proceedings of the IEEE, 1990. **78**(9): p. 1464-1480.
63. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.
64. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3): p. 1-27.
65. Suykens, J.A.K. and J. Vandewalle, *Least Squares Support Vector Machine Classifiers.* Neural Processing Letters, 1999. **9**(3): p. 293-300.
66. Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural networks, 2015. **61**: p. 85-117.
67. Zhang, L., et al., *A review on deep learning applications in prognostics and health management.* IEEE Access, 2019. **7**: p. 162415-162438.
68. Jozefowicz, R., W. Zaremba, and I. Sutskever. *An empirical exploration of recurrent network architectures*. in *International conference on machine learning*. 2015.
69. Donoghue, C., et al. *Manifold learning for automatically predicting articular cartilage morphology in the knee with data from the osteoarthritis initiative (OAI)*. in *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*. 2011.
70. Marques, J., et al., *Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from MRI.* Magn Reson Med, 2013. **70**(2): p. 568-75.
71. Du, Y., J. Shan, and M. Zhang. *Knee osteoarthritis prediction on MR images using cartilage damage index and machine learning methods*. in *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*. 2017.
72. Ashinsky, B.G., et al., *Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative.* J Orthop Res, 2017. **35**(10): p. 2243-2250.
73. Du, Y., et al., *A Novel Method to Predict Knee Osteoarthritis Progression on MRI Using Machine Learning Methods.* IEEE Trans Nanobioscience, 2018.
74. Abedin, J., et al., *Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images.* Scientific reports, 2019. **9**(1): p. 5761.
75. Pedoia, V., et al., *MRI and biomechanics multidimensional data analysis reveals R2 -R1rho as an early predictor of cartilage lesion progression in knee osteoarthritis.* J Magn Reson Imaging, 2018. **47**(1): p. 78-90.
76. Widera, P., et al., *Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data.* arXiv preprint arXiv:1909.13408, 2019.
77. Nelson, A., et al., *A machine learning approach to knee osteoarthritis phenotyping: data from the FNIH Biomarkers Consortium.* Osteoarthritis and cartilage, 2019. **27**(7): p. 994-1001.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

78. Tiulpin, A., et al., *Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data.* Scientific Reports, 2019. **9**(1): p. 1-11.
79. Halilaj, E., et al., *Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative.* Osteoarthritis and Cartilage, 2018. **26**(12): p. 1643-1650.
80. Yoo, T.K., et al., *Interpretation of movement during stair ascent for predicting severity and prognosis of knee osteoarthritis in elderly women using support vector machine.* Conf Proc IEEE Eng Med Biol Soc, 2013. **2013**: p. 192-6.
81. Lazzarini, N., et al., *A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women.* Osteoarthritis Cartilage, 2017. **25**(12): p. 2014-2021.
82. Beynon, M.J., L. Jones, and C.A. Holt, *Classification of osteoarthritic and normal knee function using three-dimensional motion analysis and the Dempster-Shafer theory of evidence.* IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans, 2006. **36**(1): p. 173-186.
83. Mezghani, N., et al., *Hierarchical analysis and classification of asymptomatic and knee osteoarthritis gait patterns using a wavelet representation of kinetic data and the nearest neighbor classifier.* Journal of Mechanics in Medicine and Biology, 2008. **8**(1): p. 45-54.
84. Mezghani, N., et al., *Automatic classification of asymptomatic and osteoarthritis knee gait patterns using kinematic data features and the nearest neighbor classifier.* IEEE Trans Biomed Eng, 2008. **55**(3): p. 1230-2.
85. Moustakidis, S.P., J.B. Theocharis, and G. Giakas, *A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements.* Med Eng Phys, 2010. **32**(10): p. 1145-60.
86. Şen Köktaş, N., et al., *A multi-classifier for grading knee osteoarthritis using gait analysis.* Pattern Recognition Letters, 2010. **31**(9): p. 898-904.
87. Kotti, M., et al., *The complexity of human walking: a knee osteoarthritis study.* PLoS One, 2014. **9**(9): p. e107325.
88. Deluzio, K.J. and J.L. Astephen, *Biomechanical features of gait waveform data associated with knee osteoarthritis: An application of principal component analysis.* Gait Posture, 2007. **25**: p. 86–93.
89. Jones, L., C.A. Holt, and M.J. Beynon, *Reduction, classification and ranking of motion analysis data: an application to osteoarthritic and normal knee function data.* Comput Methods Biomech Biomed Engin, 2008. **11**(1): p. 31-40.
90. Lim, J., J. Kim, and S. Cheon, *A Deep Neural Network-Based Method for Early Detection of Osteoarthritis Using Statistical Data.* International journal of environmental research and public health, 2019. **16**(7): p. 1281.
91. Phinyomark, A., et al., *Gender differences in gait kinematics for patients with knee osteoarthritis.* BMC Musculoskelet Disord, 2016. **17**: p. 157.
92. Moustakidis, S., et al., *Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective.* Quantum Machine Intelligence, 2019.
93. de Dieu Uwisengeyimana, J. and T. Ibrikci, *Diagnosing Knee Osteoarthritis Using Artificial Neural Networks and Deep Learning.* Biomedical Statistics and Informatics, 2017. **2**(3): p. 95.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

94. Kotti, M., et al., *Detecting knee osteoarthritis and its discriminating parameters using random forests.* Med Eng Phys, 2017. **43**: p. 19-29.
95. Yoo, T.K., et al., *Simple Scoring System and Artificial Neural Network for Knee Osteoarthritis Risk Prediction: A Cross-Sectional Study.* PLoS One, 2016. **11**(2): p. e0148724.
96. Aksehirli, Ö., et al., *Knee Osteoarthritis Diagnosis Using Support Vector Machine and Probabilistic Neural Network*. 2013.
97. Şen Köktaş, N., N. Yalabik, and G. Yavuzer. *Ensemble classifiers for medical diagnosis of knee osteoarthritis using gait data*. in *Proceedings - 5th International Conference on Machine Learning and Applications, ICMLA 2006*. 2006.
98. Long, M.J., et al., *Predicting knee osteoarthritis risk in injured populations.* Clinical Biomechanics, 2017. **47**: p. 87-95.
99. McBride, J., et al. *Neural network analysis of gait biomechanical data for classification of knee osteoarthritis*. in *Proceedings of the 2011 Biomedical Sciences and Engineering Conference: Image Informatics and Analytics in Biomedicine, BSEC 2011*. 2011.
100. Mezghani, N., et al., *Mechanical biomarkers of medial compartment knee osteoarthritis diagnosis and severity grading: Discovery phase. J. Biomech.* 2017. **52**: p. 106–112.
101. Bien, N., et al., *Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet.* PLoS medicine, 2018. **15**(11): p. e1002699.
102. En, C.Z. and T.T. Swee, *Computer-Aided Knee Osteoarthritis Classification System Using Artificial Neural Netwrok (ANN).* Journal of Medical Imaging and Health Informatics, 2013. **3**(4): p. 561-565.
103. Pedoia, V., et al., *Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort.* Osteoarthritis and cartilage, 2019. **27**(7): p. 1002-1010.
104. Kubkaddi, S. and K. Ravikumar, *Early detection of Knee Osteoarthritis using SVM Classifier.* IJSEAT, 2017. **5**(3): p. 259-262.
105. Kumarv, A. and A.K. Jayanthy. *Classification of MRI images in 2D coronal view and measurement of articular cartilage thickness for early detection of knee osteoarthritis*. in *2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*. 2017.
106. Marques, J., L.K.H. Clemmensen, and E. Dam, *Diagnosis and prognosis of Ostheoarthritis by texture analysis using sparse linear models.* 2012.
107. Anifah, L., et al., *Osteoarthritis classification using self organizing map based on gabor kernel and contrast-limited adaptive histogram equalization.* Open Biomed Eng J, 2013. **7**: p. 18-28.
108. Antony, J., et al. *Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks*. in *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016. IEEE.
109. Anifah, L., et al. *Osteoarthritis Severity Determination using Self Organizing Map Based Gabor Kernel*. in *IOP Conference Series: Materials Science and Engineering*. 2018.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

110. Minciullo, L., et al., *Indecisive trees for classification and prediction of knee osteoarthritis*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2017. p. 283-290.
111. Minciullo, L. and T. Cootes. *Fully automated shape analysis for detection of Osteoarthritis from lateral knee radiographs*. in *Proceedings - International Conference on Pattern Recognition*. 2017.
112. Bayramoglu, N., et al., *Adaptive Segmentation of Knee Radiographs for Selecting the Optimal ROI in Texture Analysis.* arXiv preprint arXiv:1908.07736, 2019.
113. Tiulpin, A. and S. Saarakkala, *Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks.* arXiv preprint arXiv:1907.08020, 2019.
114. Chen, P., et al., *Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss.* Computerized Medical Imaging and Graphics, 2019. **75**: p. 84-92.
115. Tiulpin, A., et al., *Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach.* Sci Rep, 2018. **8**(1): p. 1727.
116. Gornale, S.S., et al., *Determination of Osteoarthritis Using Histogram of Oriented Gradients and Multiclass SVM.* International Journal of Image, Graphics & Signal Processing, 2017. **9**(12).
117. Navale, D.I., R.S. Hegadi, and N. Mendgudli. *Block based texture analysis approach for knee osteoarthritis identification using SVM*. in *2015 IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE 2015*. 2016.
118. Sharma, S., S.S. Virk, and V. Jain. *Detection of osteoarthritis using SVM classifications*. in *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*. 2016.
119. Wahyuningrum, R.T., et al. *A novel hybrid of S2DPCA and SVM for knee osteoarthritis classification*. in *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA 2016 - Proceedings*. 2016.
120. Wahyuningrum, R.T., et al. *A New Approach to Classify Knee Osteoarthritis Severity from Radiographic Images based on CNN-LSTM Method*. in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. 2019. IEEE.
121. Antony, J., et al., *Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2017. p. 376-390.
122. Górriz, M., et al., *Assessing Knee OA Severity with CNN attention-based end-to-end architectures.* arXiv preprint arXiv:1908.08856, 2019.
123. Liu, B., J. Luo, and H. Huang, *Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN.* International Journal of Computer Assisted Radiology and Surgery: p. 1-10.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

124. von Tycowicz, C., *Towards Shape-based Knee Osteoarthritis Classification using Graph Convolutional Networks.* arXiv preprint arXiv:1910.06119, 2019.
125. Levinger, P., et al., *The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters.* Gait and Posture, 2009. **29**(1): p. 91-96.
126. Wittevrongel, B., et al., *Predicting Gait Retraining Strategies for Knee Osteoarthritis*. 2015.
127. Chen, H.P., et al. *Online segmentation with multi-layer SVM for knee osteoarthritis rehabilitation monitoring*. in *BSN 2016 - 13th Annual Body Sensor Networks Conference*. 2016.
128. Huang, P.C., et al. *Human motion identification for rehabilitation exercise assessment of knee osteoarthritis*. in *Proceedings of the 2017 IEEE International Conference on Applied System Innovation: Applied System Innovation for Modern Technology, ICASI 2017*. 2017.
129. Gan, H.S., et al. *Flexible non cartilage seeds for osteoarthritic magnetic resoance image of knee: Data from the osteoarthritis initiative*. in *IECBES 2016 - IEEE-EMBS Conference on Biomedical Engineering and Sciences*. 2017.
130. Ambellan, F., et al., *Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative.* Medical image analysis, 2019. **52**: p. 109-118.
131. Tack, A., A. Mukhopadhyay, and S. Zachow, *Knee menisci segmentation using convolutional neural networks: data from the osteoarthritis initiative.* Osteoarthritis and cartilage, 2018. **26**(5): p. 680-688.
132. Tack, A. and S. Zachow. *Accurate automated volumetry of cartilage of the knee using convolutional neural networks: data from the osteoarthritis initiative*. in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019. IEEE.
133. Panfilov, E., et al. *Improving Robustness of Deep Learning Based Knee MRI Segmentation: Mixup and Adversarial Domain Adaptation*. in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.
134. Tiulpin, A., I. Melekhov, and S. Saarakkala. *KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks*. in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.
135. Tiulpin, A., et al. *A novel method for automatic localization of joint area on knee plain radiographs*. in *Scandinavian Conference on Image Analysis*. 2017. Springer.
136. Gornale, S.S., et al., *Study of Segmentation Techniques for Assessment of Osteoarthritis in Knee X-ray Images.* International Journal of Image, Graphics and Signal Processing (IJIGSP), 2019. **11**(2): p. 48-57.
137. Marstal, K., et al. *Semi-automatic segmentation of knee osteoarthritic cartilage in magnetic resonance images*. in *Proceedings Elmar - International Symposium Electronics in Marine*. 2011.
138. Kashyap, S., et al., *Automated Segmentation of Knee MRI Using Hierarchical Classifiers and Just Enough Interaction Based Learning: Data from Osteoarthritis Initiative.* Med Image Comput Comput Assist Interv, 2016. **9901**: p. 344-351.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

139. Kashyap, S., et al., *Learning-Based Cost Functions for 3-D and 4-D Multi-Surface Multi-Object Segmentation of Knee MRI: Data From the Osteoarthritis Initiative.* IEEE Trans Med Imaging, 2018. **37**(5): p. 1103-1113.
140. Ababneh, S.Y. and M.N. Gurcan *An automated content-based segmentation framework: Application to MR images of knee for osteoarthritis research*. 2010 IEEE International Conference on Electro/Information Technology, EIT2010, 2010. DOI: 10.1109/EIT.2010.5612188.
141. Park, S.H., et al. *Fully automatic 3-D segmentation of knee bone compartments by iterative local branch-and-mincut on MR images from Osteoarthritis Initiative (OAI)*. in *Proceedings - International Conference on Image Processing, ICIP*. 2009.
142. Swanson, M.S., et al., *Semi-automated segmentation to assess the lateral meniscus in normal and osteoarthritic knees.* Osteoarthritis Cartilage, 2010. **18**(3): p. 344-53.
143. Tamez-Pena, J.G., et al., *Unsupervised segmentation and quantification of anatomical knee features: data from the Osteoarthritis Initiative.* IEEE Trans Biomed Eng, 2012. **59**(4): p. 1177-86.
144. Ackerman, I.N., et al., *Hip and Knee Osteoarthritis Affects Younger People, Too.* J Orthop Sports Phys Ther, 2017. **47**(2): p. 67-79.
145. Alexos, A., et al. *Physical Activity as a Risk Factor in the Progression of Osteoarthritis: A Machine Learning Perspective*. in *International Conference on Learning and Intelligent Optimization*. 2020. Springer.
146. Juszczak, P., D. Tax, and R.P. Duin. *Feature scaling in support vector data description*. in *Proc. asci*. 2002. Citeseer.
147. Dodge, Y. and D. Commenges, *The Oxford dictionary of statistical terms*. 2006: Oxford University Press on Demand.
148. Biesiada, J. and W. Duch, *Feature selection for high-dimensional data—a Pearson redundancy based filter*, in *Computer recognition systems 2*. 2007, Springer. p. 242-249.
149. Thaseen, I.S. and C.A. Kumar, *Intrusion detection model using fusion of chi-square feature selection and multi class SVM.* Journal of King Saud University-Computer and Information Sciences, 2017. **29**(4): p. 462-472.
150. Xiong, M., X. Fang, and J. Zhao, *Biomarker identification by feature wrappers.* Genome Research, 2001. **11**(11): p. 1878-1887.
151. Nie, F., et al. *Efficient and robust feature selection via joint ℓ2, 1-norms minimization*. in *Advances in neural information processing systems*. 2010.
152. Zhou, Q., H. Zhou, and T. Li, *Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features.* Knowledge-based systems, 2016. **95**: p. 1-11.
153. Al Daoud, E., *Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset.* International Journal of Computer and Information Engineering, 2019. **13**(1): p. 6-10.
154. Rockel, J.S., et al., *A classification modeling approach for determining metabolite signatures in osteoarthritis.* PloS one, 2018. **13**(6).
155. Kobayashi, T., et al., *Predictors affecting balance performances in patients with knee osteoarthritis using decision tree analysis.* Osteoarthritis and Cartilage, 2019. **27**: p. S243.
156. Peterson, L.E., *K-nearest neighbor.* Scholarpedia, 2009. **4**(2): p. 1883.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

157. Torlay, L., et al., *Machine learning–XGBoost analysis of language networks to classify patients with epilepsy.* Brain informatics, 2017. **4**(3): p. 159-169.
158. Antony, B., et al., *Do early life factors affect the development of knee osteoarthritis in later life: a narrative review.* Arthritis research & therapy, 2016. **18**(1): p. 202.
159. Toivanen, A.T., et al., *Obesity, physically demanding work and traumatic knee injury are major risk factors for knee osteoarthritis—a population-based study with a follow-up of 22 years.* Rheumatology, 2010. **49**(2): p. 308-314.
160. Jack Farr, I., L.E. Miller, and J.E. Block, *Quality of life in patients with knee osteoarthritis: a commentary on nonsurgical and surgical treatments.* The open orthopaedics journal, 2013. **7**: p. 619.
161. Ntakolia, C., et al. *A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients*. in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. 2020. IEEE.
162. Moustakidis, S., et al., *Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness.* Neural Computing and Applications, 2020: p. 1-13.
163. Kokkotis, C., et al., *Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients.* Applied Sciences, 2020. **10**(19): p. 6797.
164. Alexos, A., et al. *Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative*. in *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*. 2020. IEEE.
165. Jamshidi, A., et al., *Identification of the most important features of knee osteoarthritis structural progressors using machine learning methods.* Therapeutic advances in musculoskeletal disease, 2020. **12**: p. 1759720X20933468.
166. Wang, Y., et al., *Causal Discovery in Radiographic Markers of Knee Osteoarthritis and Prediction for Knee Osteoarthritis Severity With Attention–Long Short-Term Memory.* Frontiers in Public Health, 2020. **8**: p. 845.
167. Li, G.-Z., et al., *Asymmetric bagging and feature selection for activities prediction of drug molecules.* BMC bioinformatics, 2008. **9**(S6): p. S7.
168. Fu, G.-H., et al., *Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data.* BMC bioinformatics, 2020. **21**: p. 1-14.
169. Nimankar, S.S. and D. Vora, *Designing a Model to Handle Imbalance Data Classification Using SMOTE and Optimized Classifier*, in *Data Management, Analytics and Innovation*. 2020, Springer. p. 323-334.
170. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
171. Han, H., W.-Y. Wang, and B.-H. Mao. *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning*. in *International conference on intelligent computing*. 2005. Springer.
172. Yen, S.-J. and Y.-S. Lee, *Cluster-based under-sampling approaches for imbalanced data distributions.* Expert Systems with Applications, 2009. **36**(3): p. 5718-5727.
173. He, H., et al. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. 2008. IEEE.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

174. Chawla, N.V., et al. *SMOTEBoost: Improving prediction of the minority class in boosting*. in *European conference on principles of data mining and knowledge discovery*. 2003. Springer.
175. Hanifah, F.S., H. Wijayanto, and A. Kurnia, *SMOTEBagging algorithm for imbalanced dataset in logistic regression analysis (case: credit of bank X).* Applied Mathematical Sciences, 2015. **9**(138): p. 6857-6865.
176. Elkan, C. *The foundations of cost-sensitive learning*. in *International joint conference on artificial intelligence*. 2001. Lawrence Erlbaum Associates Ltd.
177. Ling, C.X. and V.S. Sheng, *Cost-sensitive learning and the class imbalance problem*. 2008, Citeseer. p. 231-235.
178. Shin, H.J., D.-H. Eom, and S.-S. Kim, *One-class support vector machines—an application in machine fault detection and classification.* Computers & Industrial Engineering, 2005. **48**(2): p. 395-408.
179. Seo, K.-K., *An application of one-class support vector machines in content-based image retrieval.* Expert Systems with Applications, 2007. **33**(2): p. 491-498.
180. Ertekin, S., J. Huang, and C.L. Giles. *Active learning for class imbalance problem*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007.
181. Attenberg, J. and S. Ertekin, *Class imbalance and active learning.* Imbalanced Learning: Foundations, Algorithms, and Applications, 2013: p. 101-149.
182. Shanker, M., M.Y. Hu, and M.S. Hung, *Effect of data standardization on neural network training.* Omega, 1996. **24**(4): p. 385-397.
183. Lundberg, S.M. and S.-I. Lee. *A unified approach to interpreting model predictions*. in *Advances in neural information processing systems*. 2017.
184. Roffo, G., et al. *Infinite latent feature selection: A probabilistic latent graph-based ranking approach*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
185. Roffo, G., S. Melzi, and M. Cristani. *Infinite feature selection*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
186. Shahbaz, M.B., et al. *On efficiency enhancement of the correlation-based feature selection for intrusion detection systems*. in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. 2016. IEEE.
187. Hagos, D.H., et al. *Enhancing security attacks analysis using regularized machine learning techniques*. in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. 2017. IEEE.
188. Nguyen, H.T., K. Franke, and S. Petrovic. *Towards a generic feature-selection measure for intrusion detection*. in *2010 20th International Conference on Pattern Recognition*. 2010. IEEE.
189. Cooper, C., et al., *Risk factors for the incidence and progression of radiographic knee osteoarthritis.* Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 2000. **43**(5): p. 995-1000.
190. Hartley, A., et al., *Individuals with High Bone Mass have increased progression of radiographic and clinical features of knee osteoarthritis.* Osteoarthritis and Cartilage, 2020.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

191. Blagojevic, M., et al., *Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis.* Osteoarthritis and cartilage, 2010. **18**(1): p. 24-33.
192. Heidari, B., *Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I.* Caspian journal of internal medicine, 2011. **2**(2): p. 205.
193. Santos, M.S., et al., *Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier].* ieee ComputatioNal iNtelligeNCe magaziNe, 2018. **13**(4): p. 59-76.
194. Malanga, G., et al., *Knee Osteoarthritis Treatment Costs in the Medicare Patient Population.* American health & drug benefits, 2020. **13**(4): p. 144.
195. Johnson, V.L. and D.J. Hunter, *The epidemiology of osteoarthritis.* Best practice & research Clinical rheumatology, 2014. **28**(1): p. 5-15.
196. Mahir, L., et al., *Impact of knee osteoarthritis on the quality of life.* Annals of physical and rehabilitation medicine, 2016. **59**: p. e159.
197. Christodoulou, E., et al. *Exploring deep learning capabilities in knee osteoarthritis case study for classification*. in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. 2019. IEEE.
198. Kwon, S.B., et al., *Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images.* IEEE Access, 2020. **8**: p. 120597-120603.
199. Hardi, S. and A. Triwiyono. *Expert System for Diagnosing Osteoarthritis with Fuzzy Tsukamoto Method*. in *Journal of Physics: Conference Series*. 2020. IOP Publishing.
200. Pal, J.K., S.S. Ray, and S.K. Pal, *Fuzzy mutual information based grouping and new fitness function for PSO in selection of miRNAs in cancer.* Computers in biology and medicine, 2017. **89**: p. 540-548.
201. Dai, J. and J. Chen, *Feature selection via normative fuzzy information weight with application into tumor classification.* Applied Soft Computing, 2020. **92**: p. 106299.
202. Lin, Y., et al., *Streaming feature selection for multilabel learning based on fuzzy mutual information.* IEEE Transactions on Fuzzy Systems, 2017. **25**(6): p. 1491-1507.
203. Jaganathan, P. and R. Kuppuchamy, *A threshold fuzzy entropy based feature selection for medical database classification.* Computers in Biology and Medicine, 2013. **43**(12): p. 2222-2229.
204. Wang, C., et al., *A fitting model for feature selection with fuzzy rough sets.* IEEE Transactions on Fuzzy Systems, 2016. **25**(4): p. 741-753.
205. Qian, Y., et al., *Fuzzy-rough feature selection accelerator.* Fuzzy Sets and Systems, 2015. **258**: p. 61-78.
206. Silva-Ramírez, E.-L., et al., *Missing value imputation on missing completely at random data using multilayer perceptrons.* Neural Networks, 2011. **24**(1): p. 121-129.
207. Pihera, J. and N. Musliu. *Application of machine learning to algorithm selection for TSP*. in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. 2014. IEEE.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

208. Japkowicz, N. *Learning from imbalanced data sets: a comparison of various strategies*. in *AAAI workshop on learning from imbalanced data sets*. 2000. AAAI Press Menlo Park, CA.
209. Vergara, J.R. and P.A. Estévez, *A review of feature selection methods based on mutual information.* Neural computing and applications, 2014. **24**(1): p. 175-186.
210. Suchwalko, A., I. Buzalewicz, and H. Podbielska. *Identification of bacteria species by using morphological and textural properties of bacterial colonies diffraction patterns*. in *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*. 2013. International Society for Optics and Photonics.
211. Zhao, J., et al. *FeatureExplorer: Interactive feature selection and exploration of regression models for hyperspectral images*. in *2019 IEEE Visualization Conference (VIS)*. 2019. IEEE.
212. Ding, J., J. Shi, and F.-X. Wu, *SVM-RFE based feature selection for tandem mass spectrum quality assessment.* International journal of data mining and bioinformatics, 2011. **5**(1): p. 73-88.
213. Ye, Y., et al. *Optimal feature selection for EMG-based finger force estimation using lightGBM model*. in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2019. IEEE.
214. Mate, Y. and N. Somai. *Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction*. in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2021. IEEE.
215. Gayathri, B. and C. Sumathi. *Mamdani fuzzy inference system for breast cancer risk detection*. in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. 2015. IEEE.
216. Ram, M., A. Najafi, and M.T. Shakeri, *Classification and biomarker genes selection for cancer gene expression data using random forest.* Iranian journal of pathology, 2017. **12**(4): p. 339.
217. Parisi, L., et al. *A novel comparison of artificial intelligence methods for diagnosing knee osteoarthritis*. in *XXV congress of the international society of biomechanics*. 2015.
218. Long, N.P., et al., *Efficacy of integrating a novel 16-gene biomarker panel and intelligence classifiers for differential diagnosis of rheumatoid arthritis and osteoarthritis.* Journal of clinical medicine, 2019. **8**(1): p. 50.
219. Wong, T.-T. and P.-Y. Yeh, *Reliable accuracy estimates from k-fold cross validation.* IEEE Transactions on Knowledge and Data Engineering, 2019. **32**(8): p. 1586-1594.
220. Ghosh, M. and G. Sanyal, *An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning.* Journal of Big Data, 2018. **5**(1): p. 1-25.
221. Nohara, Y., et al. *Explanation of machine learning models using improved Shapley Additive Explanation*. in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019.
222. Ntakolia, C., et al., *Prediction of Joint Space Narrowing Progression in Knee Osteoarthritis Patients.* Diagnostics, 2021. **11**(2): p. 285.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

223. Katz, J.N., K.R. Arant, and R.F. Loeser, *Diagnosis and treatment of hip and knee osteoarthritis: a review.* Jama, 2021. **325**(6): p. 568-578.
224. Roos, E.M. and L.S. Lohmander, *The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis.* Health and quality of life outcomes, 2003. **1**(1): p. 1-8.
225. Prodromos, C.C., et al., *A meta-analysis of the incidence of anterior cruciate ligament tears as a function of gender, sport, and a knee injury-reduction regimen.* Arthroscopy, 2007. **23**(12): p. 1320-1325.e6.
226. Moses, B., J. Orchard, and J. Orchard, *Systematic review: Annual incidence of ACL injury and surgery in various populations.* Res Sports Med, 2012. **20**(3-4): p. 157-79.
227. Kanamori, A., et al., *The effect of axial tibial torque on the function of the anterior cruciate ligament: a biomechanical study of a simulated pivot shift test.* Arthroscopy, 2002. **18**(4): p. 394-8.
228. Zantop, T., et al., *The role of the anteromedial and posterolateral bundles of the anterior cruciate ligament in anterior tibial translation and internal rotation.* Am J Sports Med, 2007. **35**(2): p. 223-7.
229. Hanzlíková, I., et al., *The effect of proprioceptive knee bracing on knee stability during three different sport related movement tasks in healthy subjects and the implications to the management of Anterior Cruciate Ligament (ACL) injuries.* Gait & posture, 2016. **48**: p. 165-170.
230. Rezende, F.C., et al., *Does combined intra-and extraarticular ACL reconstruction improve function and stability? A meta-analysis.* Clinical Orthopaedics and Related Research®, 2015. **473**(8): p. 2609-2618.
231. Andriacchi, T.P. and C.O. Dyrby, *Interactions between kinematics and loading during walking for the normal and ACL deficient knee.* J Biomech, 2005. **38**(2): p. 293-8.
232. Georgoulis, A.D., et al., *Three-dimensional tibiofemoral kinematics of the anterior cruciate ligament-deficient and reconstructed knee during walking.* Am J Sports Med, 2003. **31**(1): p. 75-9.
233. Tsarouhas, A., et al., *Three-Dimensional Kinematic and Kinetic Analysis of Knee Rotational Stability After Single- and Double-Bundle Anterior Cruciate Ligament Reconstruction.* Arthroscopy - Journal of Arthroscopic and Related Surgery, 2010. **26**(7): p. 885-893.
234. Moore, J.M., et al., *Lower limb biomechanics before and after anterior cruciate ligament reconstruction: A systematic review.* Journal of Biomechanics, 2020. **106**: p. 109828.
235. Defrate, L.E., et al., *The 6 degrees of freedom kinematics of the knee after anterior cruciate ligament deficiency: an in vivo imaging analysis.* Am J Sports Med, 2006. **34**(8): p. 1240-6.
236. Andriacchi, T.P., S. Koo, and S.F. Scanlan, *Gait mechanics influence healthy cartilage morphology and osteoarthritis of the knee.* J Bone Joint Surg Am, 2009. **91 Suppl 1**(Suppl 1): p. 95-101.
237. Andriacchi, T.P., et al., *Rotational changes at the knee after ACL injury cause cartilage thinning.* Clin Orthop Relat Res, 2006. **442**: p. 39-44.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

238. Chaudhari, A.M., et al., *Knee kinematics, cartilage morphology, and osteoarthritis after ACL injury.* Med Sci Sports Exerc, 2008. **40**(2): p. 215-22.
239. Butler, R.J., et al., *Gait mechanics after ACL reconstruction: implications for the early onset of knee osteoarthritis.* Br J Sports Med, 2009. **43**(5): p. 366-70.
240. Mall, N.A., et al., *Incidence and trends of anterior cruciate ligament reconstruction in the United States.* Am J Sports Med, 2014. **42**(10): p. 2363-70.
241. Slater, L.V., et al., *Progressive Changes in Walking Kinematics and Kinetics After Anterior Cruciate Ligament Injury and Reconstruction: A Review and Meta-Analysis.* J Athl Train, 2017. **52**(9): p. 847-860.
242. Papannagari, R., et al., *In vivo kinematics of the knee after anterior cruciate ligament reconstruction: a clinical and functional evaluation.* Am J Sports Med, 2006. **34**(12): p. 2006-12.
243. Andriacchi, T.P. and A. Mündermann, *The role of ambulatory mechanics in the initiation and progression of knee osteoarthritis.* Curr Opin Rheumatol, 2006. **18**(5): p. 514-8.
244. Hurwitz, D.E., et al., *Knee pain and joint loading in subjects with osteoarthritis of the knee.* J Orthop Res, 2000. **18**(4): p. 572-9.
245. Sharma, L., et al., *Knee adduction moment, serum hyaluronan level, and disease severity in medial tibiofemoral osteoarthritis.* Arthritis Rheum, 1998. **41**(7): p. 1233-40.
246. Di Stasi, S.L., et al., *Gait patterns differ between ACL-reconstructed athletes who pass return-to-sport criteria and those who fail.* Am J Sports Med, 2013. **41**(6): p. 1310-8.
247. Timoney, J.M., et al., *Return of normal gait patterns after anterior cruciate ligament reconstruction.* Am J Sports Med, 1993. **21**(6): p. 887-9.
248. Shin, C.S., et al., *Influence of patellar ligament insertion angle on quadriceps usage during walking in anterior cruciate ligament reconstructed subjects.* J Orthop Res, 2009. **27**(6): p. 730-5.
249. Bayliss, L. and L.D. Jones, *The role of artificial intelligence and machine learning in predicting orthopaedic outcomes.* Bone Joint J, 2019. **101-b**(12): p. 1476-1478.
250. Sanchez-Santos, M.T., et al., *Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery.* Scientific reports, 2018. **8**(1): p. 3381-3381.
251. Olczak, J., et al., *Artificial intelligence for analyzing orthopedic trauma radiographs.* Acta Orthop, 2017. **88**(6): p. 581-586.
252. Kunze, K.N., et al., *Development of Machine Learning Algorithms to Predict Patient Dissatisfaction After Primary Total Knee Arthroplasty.* The Journal of Arthroplasty, 2020. **35**(11): p. 3117-3122.
253. Brisson, N.M., et al., *Association of machine learning based predictions of medial knee contact force with cartilage loss over 2.5 years in knee osteoarthritis.* Arthritis Rheumatol, 2021.
254. Moustakidis, S.P., J.B. Theocharis, and G. Giakas, *Feature selection based on a fuzzy complementary criterion: application to gait recognition using ground reaction forces.* Computer Methods in Biomechanics and Biomedical Engineering, 2012. **15**(6): p. 627-644.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

255. Mazlan, S., M.Z. Ayob, and Z. Bakti, *Anterior cruciate ligament (ACL) injury classification system using support vector machine (SVM).* 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), 2017: p. 1-5.
256. Chang, P.D., T.T. Wong, and M.J. Rasiej, *Deep Learning for Detection of Complete Anterior Cruciate Ligament Tear.* J Digit Imaging, 2019. **32**(6): p. 980-986.
257. Christian, J., et al., *Computer aided analysis of gait patterns in patients with acute anterior cruciate ligament injury.* Clinical Biomechanics, 2016. **33**: p. 55-60.
258. Zeng, W., S.A. Ismail, and E. Pappas, *Detecting the presence of anterior cruciate ligament injury based on gait dynamics disparity and neural networks.* Artificial Intelligence Review, 2020. **53**(5): p. 3153-3176.
259. Tedesco, S., et al., *Motion Sensors-Based Machine Learning Approach for the Identification of Anterior Cruciate Ligament Gait Patterns in On-the-Field Activities in Rugby Players.* Sensors, 2020. **20**(11): p. 3029.
260. Tsatalas, T., et al., *Walking kinematics and kinetics following eccentric exercise-induced muscle damage.* Journal of Electromyography and Kinesiology, 2013. **23**(5): p. 1229-1236.
261. Tsatalas, T., et al., *Altered Drop Jump Landing Biomechanics Following Eccentric Exercise-Induced Muscle Damage.* Sports, 2021. **9**(2): p. 24.
262. Ehrig, R.M., et al., *A survey of formal methods for determining the centre of rotation of ball joints.* Journal of Biomechanics, 2006. **39**(15): p. 2798-2809.
263. Ehrig, R.M., et al., *A survey of formal methods for determining functional joint axes.* Journal of Biomechanics, 2007. **40**(10): p. 2150-2157.
264. Urbanowicz, R.J., et al., *Relief-based feature selection: Introduction and review.* Journal of biomedical informatics, 2018. **85**: p. 189-203.
265. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* Journal of chemical information and computer sciences, 2003. **43**(6): p. 1947-1958.
266. Podgorelec, V., et al., *Decision trees: an overview and their use in medicine.* Journal of medical systems, 2002. **26**(5): p. 445-463.
267. Miranda, E., et al., *Detection of cardiovascular disease risk's level for adults using naive Bayes classifier.* Healthcare informatics research, 2016. **22**(3): p. 196-205.
268. Ma, Y., et al., *Classification of motor imagery EEG signals with support vector machines and particle swarm optimization.* Computational and mathematical methods in medicine, 2016. **2016**.
269. Subasi, A. and E. Ercelebi, *Classification of EEG signals using neural network and logistic regression.* Computer methods and programs in biomedicine, 2005. **78**(2): p. 87-99.
270. Park, J. and D.H. Lee, *Privacy preserving k-nearest neighbor for medical diagnosis in e-health cloud.* Journal of healthcare engineering, 2018. **2018**.
271. Parsa, A.B., et al., *Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis.* Accident Analysis & Prevention, 2020. **136**: p. 105405.
272. Bewick, V., L. Cheek, and J. Ball, *Statistics review 9: one-way analysis of variance.* Critical care, 2004. **8**(2): p. 1-7.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

273. Gerald, B., *A brief review of independent, dependent and one sample t-test.* International Journal of Applied Mathematics and Theoretical Physics, 2018. **4**(2): p. 50-54.
274. Knoll, Z., L. Kocsis, and R.M. Kiss, *Gait patterns before and after anterior cruciate ligament reconstruction.* Knee Surg Sports Traumatol Arthrosc, 2004. **12**(1): p. 7-14.
275. Asaeda, M., et al., *Gender differences in the restoration of knee joint biomechanics during gait after anterior cruciate ligament reconstruction.* The Knee, 2017. **24**(2): p. 280-288.
276. Dikopoulou, Z., S. Moustakidis, and P. Karlsson *GLIME: A new graphical methodology for interpretable model-agnostic explanations*. 2021. arXiv:2107.09927.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Appendixes

Appendix A

Table A. Selected features that led to the overall best Knee Osteoarthritis (KOA) prediction performance in our study.

Feature	Description	Category
P02WTGA	Above weight cut-off for age/gender group (calc, used for study eligibility)	Subject characteristics
V00WPRKN2	Right knee pain: stairs, last 7 days	Symptoms
V00RXANALG	Rx Analgesic use indicator (calc)	Medical history
V00PCTSMAL	Block Brief 2000: error flag, percent of foods marked as small portion (calc)	Nutrition
V00GLUC	Used glucosamine for joint pain or arthritis, past 6 months	Medical history
V00GLCFQCV	Glucosamine frequency of use, past 6 months (calc)	Medical history
V00CHON	Used chondroitin sulfate for joint pain or arthritis, past 6 months	Medical history
V00CHNFQCV	Chondroitin sulfate frequency of use, past 6 months (calc)	Medical history
V00BAPCARB	Block Brief 2000: daily % of calories from carbohydrate, alcoholic beverages excluded from denominator (kcal) (calc)	Nutrition
P02KPNRCV	Right knee pain, aching or stiffness: more than half the days of a month, past 12 months (calc, used for study eligibility)	Symptoms
P01XRKOA	Baseline radiographic knee OA status by person (calc)	Medical imaging outcome
P01SVLKOST	Left knee baseline x-ray: evidence of knee osteophytes (calc)	Medical imaging outcome
P01OAGRDL	Left knee baseline x-ray: composite OA grade (calc)	Medical imaging outcome
P01GOUTCV	Doctor said you had gout (calc)	Medical history
V00WTMAXKG	Maximum adult weight, self-reported (kg) (calc)	Subject characteristics
V00WSRKN1	Right knee stiffness: in morning, last 7 days	Symptoms
V00WOMSTFR	Right knee: WOMAC Stiffness Score (calc)	Symptoms
V00SF1	In general, how is health	Behavioural
V00RKMTTPN	Right knee exam: medial tibiofemoral pain/tenderness present on exam	Physical exam
V00RFXCOMP	Isometric strength: right knee flexion, able to complete (3) measurements	Physical exam

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

V00PCTFAT	Block Brief 2000: daily percent of calories from fat (kcal) (calc)	Nutrition
V00PCTCARB	Block Brief 2000: daily percent of calories from carbohydrate (kcal) (calc)	Nutrition
V00PASE	Physical Activity Scale for the Elderly (PASE) score (calc)	Physical activity
V00LUNG	Charlson Comorbidity: have emphysema, chronic bronchitis or chronic obstructive lung disease (also called COPD)	Medical history
V00KSXLKN1	Left knee symptoms: swelling, last 7 days	Symptoms
V00FFQSZ16	Block Brief 2000: rice/dishes made with rice, how much each time	Nutrition
V00FFQSZ14	Block Brief 2000: white potatoes not fried, how much each time	Nutrition
V00FFQSZ13	Block Brief 2000: french fries/fried potatoes/hash browns, how much each time	Nutrition
V00FFQ69	Block Brief 2000: regular soft drinks/bottled drinks like Snapple (not diet drinks), drink how often, past 12 months	Nutrition
V00FFQ59	Block Brief 2000: ice cream/frozen yogurt/ice cream bars, eat how often, past 12 months	Nutrition
V00FFQ37	Block Brief 2000: fried chicken, at home or in a restaurant, eat how often, past 12 months	Nutrition
V00DTCAFFN	Block Brief 2000: daily nutrients from food, caffeine (mg) (calc)	Nutrition
V00DILKN11	Left knee difficulty: socks off, last 7 days	Symptoms
V00CESD13	CES-D: how often talked less than usual, past week	Behavioural
V00ABCIRC	Abdominal circumference (cm) (calc)	Subject characteristics
TIMET1	20-m walk: trial 1 time to complete (sec.hundredths/sec)	Physical exam
STEPST1	20-m walk: trial 1 number of steps	Physical exam
PASE6	Leisure activities: muscle strength/endurance, past 7 days	Physical activity
P02KPNLCV	Left knee pain, aching or stiffness: more than half the days of a month, past 12 months (calc, used for study eligibility)	Symptoms
P01WEIGHT	Average current scale weight (kg) (calc)	Subject characteristics
P01SVRKOST	Right knee baseline x-ray: evidence of knee osteophytes (calc)	Medical imaging outcome
P01SVRKJSL	Right knee baseline x-ray: evidence of knee lateral joint space narrowing (calc)	Medical imaging outcome
P01RXRKOA2	Right knee baseline x-ray: osteophytes and JSN (calc)	Medical imaging outcome

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

P01RXRKOA	Right knee baseline radiographic OA (definite osteophytes, calc, used in OAI definition of symptomatic knee OA)	Medical imaging outcome
P01RSXKOA	Right knee baseline symptomatic OA status (calc)	Medical imaging outcome
P01OAGRDR	Right knee baseline x-ray: composite OA grade (calc)	Medical imaging outcome
P01LXRKOA2	Left knee baseline x-ray: osteophytes and JSN (calc)	Medical imaging outcome
P01LXRKOA	Left knee baseline radiographic OA (definite osteophytes, calc, used in OAI definition of symptomatic knee OA)	Medical imaging outcome
P01BMI	Body mass index (calc)	Subject characteristics
P01ARTDRCV	Seeing doctor/other professional for knee arthritis (calc)	Medical history
KSXRKN2	Right knee symptoms: feel grinding, hear clicking or any other type of noise when knee moves, last 7 days	Symptoms
KPRKN1	Right knee pain: twisting/pivoting on knee, last 7 days	Symptoms
DIRKN7	Right knee difficulty: in car/out of car, last 7 days	Symptoms
rkdefcv	Right knee exam: alignment varus or valgus (calc)	Physical exam
lkdefcv	Left knee exam: alignment varus or valgus (calc)	Physical exam

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Appendix B

Table B. Selected features that led to the best overall KOA prediction performance in our study. The features have been ranked according to their impact on the classification result as calculated by SHapley Additive exPlanations (SHAP).

Selected Features	Description	Feature Category
P01SVLKOST	Left knee baseline X-ray: evidence of knee osteophytes	Medical imaging outcome
P01BMI	Body mass index	Subject characteristics
V00SUPCA	Block Brief 2000: average daily nutrients from vitamin supplements, calcium (mg)	Nutrition
V00EDCV	Highest grade or year of school completed	Behavioral
V00FFQ59	Block Brief 2000: ice cream/frozen yogurt/ice cream bars, eat how often, past 12 months	Nutrition
V00KQOL2	Quality of life: modified lifestyle to avoid potentially damaging activities to knee(s)	Behavioral
V00CHNFQCV	Chondroitin sulfate frequency of use, past 6 months	Medical history
V00WOMSTFR	Right knee: WOMAC Stiffness Score	Symptoms
V00FFQSZ13	Block Brief 2000: french fries/fried potatoes/hash browns, how much each time	Nutrition
V00KQOL4	Quality of life: in general, how much difficulty have with knee(s)	Behavioral
P01HEIGHT	Average height (mm)	Subject characteristics
V00lfTHPL	Left Flexion MAX Force High Production Limit	Physical exam
V00rkdefcv	Right knee exam: alignment varus or valgus	Physical exam
V00FFQ19	Block Brief 2000: green beans/green peas, eat how often, past 12 months	Nutrition
V00FFQ33	Block Brief 2000: beef steaks/roasts/pot roast (including in frozen dinners/sandwiches), eat how often, past 12 months	Nutrition
KPLKN1	Left knee pain: twisting/pivoting on knee, last 7 days	Symptoms
PASE2	Leisure activities: walking, past 7 days	Physical activity
V00INCOME	Yearly income	Behavioral
V00PA130CV	How often climb up total of 10 or more flights of stairs during typical week, past 30 days	Physical activity
V00CESD9	How often thought my life had been a failure, past week	Behavioral
PASE6	Leisure activities: muscle strength/endurance, past 7 days	Physical activity
DIRKN16	Right knee difficulty: heavy chores, last 7 days	Symptoms

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

V00SUPB2	Block Brief 2000: average daily nutrients from vitamin supplements, B2 (mg)	Nutrition
STEPST1	20-meter walk: trial 1 number of steps	Physical exam
V00FFQ12	Block Brief 2000: any other fruit (e.g., grapes/melon/strawberries/peaches), eat how often, past 12 months	Nutrition
KSXRKN1	Right knee symptoms: swelling, last 7 days	Symptoms
V00lfmaxf	Left Flexion MAX Force	Physical exam
V00rfTHPL	Right Flexion MAX Force High Production Limit	Physical exam
RKALNMT	Right knee exam: alignment, degrees (valgus negative)	Physical exam
CEMPLOY	Current employment	Behavioral
V00KOOSYML	Left knee: KOOS Symptoms Score	Symptoms
V00WPLKN2	Left knee pain: stairs, last 7 days	Symptoms
V00RA	Charlson Comorbidity: have rheumatoid arthritis	Medical history
V00SUPFOL	Block Brief 2000: average daily nutrients from vitamin supplements, folate (mcg)	Nutrition
V00RXCHOND	Rx Chondroitin sulfate use indicator	Medical history

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Appendix C

Table C. The 21 most informative selected risk factors as described in OAI database.

Selected Features	Description
P02ELGRISK	Knee symptoms, risk factors, or both, status at IEI/SV
P01BMI	Body mass index
V00AGE	Age
P01WEIGHT	Average current scale weight (kg)
V00LKFHDEG	Left knee exam: flexion contracture/hyperextension, degrees (contracture positive)
V00KOOSKPR	Right knee: KOOS Pain Score
P01MOMHRCV	Mother had hip replacement surgery
P02PA1	Climb up total of 10 or more flights of stairs on most days
P01KSX	Frequent knee pain status by person
V00RKFHDEG	Right knee exam: flexion contracture/hyperextension, degrees (contracture positive)
V00WTMAXKG	Maximum adult weight, self-reported (kg)
P02KSURG	Either knee, history of knee surgery
V00lfTHRL	Left Flexion MAX Force High Relaxation Limit
V00BAPFAT	Block Brief 2000: daily % of calories from fat, alcoholic beverages excluded from denominator (kcal)
V00RPAVG	Radial pulse: average beats per minute
V00PASE	Physical Activity Scale for the Elderly (PASE) score
V00KOOSQOL	KOOS Quality of Life Score
V00LFXCOMP	Isometric strength: left knee flexion, able to complete (3) measurements
V00BPDIAS	Blood pressure: diastolic (mm Hg)
V00PA430CV	How often lift or move objects weighing 25 pounds or more by hand during a typical week, past 30 days

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

V00KPLKN1	Left knee pain: twisting/pivoting on knee, last 7 days

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Annexes

Annex A: Ethics

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΦΥΣΙΚΗΣ ΑΓΩΓΗΣ ΚΑΙ ΑΘΛΗΤΙΣΜΟΥ

Εσωτερική Επιτροπή Δεοντολογίας

Τρίκαλα: 3/6/2020
Αριθμ. Πρωτ:1660

Βεβαίωση έγκρισης της πρότασης για διεξαγωγή Έρευνας με τίτλο: Διερεύνηση και ανάπτυξη αλγορίθμων μηχανικής μάθησης για ανάλυση Εμβιομηχανικών δεδομένων με εφαρμογή στη βελτίωση της ποιότητας ζωής

Επιστημονικός υπεύθυνος – επιβλέπων: Δρ. Τσαόπουλος Δημήτριος
Ιδιότητα: Ερευνητής Β'
Τμήμα: Ινστιτούτο Βιο-οικονομίας και Αγρο-τεχνολογίας
Ίδρυμα: Εθνικό Κέντρο Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ)

Κύριος ερευνητής - φοιτητής: Κοκκότης Χρήστος
Πρόγραμμα Σπουδών: Υποψηφίου Διδάκτορα
Ίδρυμα: Πανεπιστήμιο Θεσσαλίας
Τμήμα: Τ.Ε.Φ.Α.Α

Η προτεινόμενη έρευνα θα είναι: Διδακτορική Έρευνα

Τηλ. επικοινωνίας: 6978526312
Email επικοινωνίας: chkokkotis@gmail.com

Η Εσωτερική Επιτροπή Δεοντολογίας του Τ.Ε.Φ.Α.Α., Πανεπιστημίου Θεσσαλίας μετά την υπ. Αριθμ. 3-1/3-6-2020 συνεδρίασή της εγκρίνει τη διεξαγωγή της προτεινόμενης έρευνας.

Ο Πρόεδρος της
Εσωτερικής Επιτροπής
Δεοντολογίας – ΤΕΦΑΑ

Τσιόκανος Αθανάσιος
Καθηγητής

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Annex B: Candidate's responsibilities throughout the study

During the PhD programme the candidate had the following duties:

- Completion of the courses required for the accomplishment of 10 to 25 ECTS
- Skills Development
- Submission of the required documents to receive ethical approval for the study
- Literature review
- Data collection and biomechanical analysis
- Data analysis as well as the implementation of artificial intelligence tools and statistical analysis
- Preparation of scientific manuscripts and submission in peer-review for publication
- Presentation of scientific results in journal clubs, workshops and conferences
- Writing the doctoral thesis and
- Public defense of the thesis

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Annex C: Skills acquired during the PhD programme

The candidate has gained considerable experience and advanced knowledge in the field of machine learning and biomechanics. The skills acquired are the following:

- Design and implementation of clinical trials
- Programming with Python and MATLAB languages
- Development of robust feature selection techniques and implementation of Machine Learning models for prediction and diagnosis tasks
- Use of advanced tools for interpretation and explainability of Machine Learning models
- Data collection and biomechanical analysis using 3D motion capture system (Vicon, UK)
- Participation as supporting personnel in National and European funded research projects
- Academic writing including scientific papers, research proposals and grands

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Annex D: Publications during PhD Studies

Journals

- Kokkotis, C., Moustakidis, S., Baltzopoulos, V., Giakas, G., & Tsaopoulos, D. (2021, March). Identifying robust risk factors for knee osteoarthritis progression: An evolutionary machine learning approach. In Healthcare (Vol. 9, No. 3, p. 260). Multidisciplinary Digital Publishing Institute.
- Kokkotis, C., Moustakidis, S., Giakas, G., & Tsaopoulos, D. (2020). Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients. Applied Sciences, 10(19), 6797.
- Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. E. (2020). Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open, 2(3), 100069.
- Ntakolia, C., Kokkotis, C., Moustakidis, S., & Tsaopoulos, D. (2021). Prediction of Joint Space Narrowing Progression in Knee Osteoarthritis Patients. Diagnostics, 11(2), 285.
- Hassandra, M., Galanis, E., Hatzigeorgiadis, A., Goudas, M., Mouzakidis, C., Karathanasi, E. M., ... & Theodorakis, Y. (2021). A Virtual Reality App for Physical and Cognitive Training of Older People With Mild Cognitive Impairment: Mixed Methods Feasibility Study. JMIR Serious Games, 9(1), e24170.
- Moustakidis, S., Christodoulou, E., Papageorgiou, E., Kokkotis, C., Papandrianos, N., & Tsaopoulos, D. (2019). Application of machine intelligence for osteoarthritis classification: A classical implementation and a quantum perspective. Quantum Machine Intelligence, 1(3), 73-86.
- Kokkotis, C., Ntakolia, C., Moustakidis, S., Giakas, G., & Tsaopoulos, D. (2021). Explainable Machine Learning for Knee Osteoarthritis Diagnosis Based on a Novel Fuzzy Feature Selection Methodology (Has been submitted).
- Kokkotis, C., Moustakidis, S., Tsatalas, T., Chalatsis, G., Ntakolia, C., Konstadakos, S., Hantes, M., Giakas, G., & Tsaopoulos, D. (2021). Leveraging explainable machine learning to identify gait biomechanical parameters associated with Anterior Cruciate Ligament injury (Has been submitted).
- Ntakolia, C., Kokkotis, C., Moustakidis, S., & Tsaopoulos, D. (2021). Identification of most important features based on a fuzzy ensemble technique: Evaluation on Joint Space Narrowing Progression in Knee Osteoarthritis Patients (Has been submitted).

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Chapters

- Alexos, A., Moustakidis, S., Kokkotis, C., & Tsaopoulos, D. (2020, May). Physical activity as a risk factor in the progression of osteoarthritis: a machine learning perspective. In International Conference on Learning and Intelligent Optimization (pp. 16-26). Springer, Cham.
- Moustakidis, S., Kokkotis, C., & Tsaopoulos, D. (2021). Patient Specific Modelling of Pain Progression: A Use Case on Knee Osteoarthritis Patients Using Machine Learning algorithms (Has been submitted).

Conferences

- Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. (2020, July). A Machine Learning workflow for Diagnosis of Knee Osteoarthritis with a focus on post-hoc explainability. In 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA (pp. 1-7). IEEE.
- Ntakolia, C., Kokkotis, C., Moustakidis, S., & Tsaopoulos, D. (2020, October). A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients. In 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 934-941). IEEE.
- Alexos, A., Kokkotis, C., Moustakidis, S., Papageorgiou, E., & Tsaopoulos, D. (2020, July). Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative. In 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA (pp. 1-7). IEEE.
- Piromalis, D. D., Kokkotis, C., Tsatalas, T., Bellis, G., Tsaopoulos, D., Zikos, P., ... & Papoutsidakis, M. (2021, March). Commercially Available Sensor-based Monitoring and Support Systems in Parkinson's Disease: An Overview. In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 430-438). IEEE.

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186

Annex E: Awards during PhD Studies

20th International Conference on BioInformatics & BioEngineering (BIBE)
Oct 26-28, 2020, Cincinnati, Ohio, USA

BIBE

On behalf of the IEEE Computer Society the

Best Student Paper Award

Presented to

Charis Ntakolia, Christos Kokkotis, Serafeim Moustakidis and Dimitris Tsaopoulos

In recognition for their paper *(A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients)* **in the BIBE-2020 Conference**

BIBE-2020 Steering & Awards Committee

IEEE COMPUTER SOCIETY

Institutional Repository - Library & Information Centre - University of Thessaly
06/05/2023 00:55:09 EEST - 54.91.76.186